# Die differentielle Analyse von Sorting-Daten durch hierarchische Cluster-Analysen

**DISSERTATION**

zur Erlangung des Grades eines
Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von

Helmut Roßbach
aus Hamburg

Saarbrücken 1975

# Die differentielle Analyse von Sorting-Daten durch hierarchische Cluster-Analysen

**DISSERTATION**

zur Erlangung des Grades eines
Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von

Helmut Roßbach
aus Hamburg

Saarbrücken 1975

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41547310_0032

# DIE DIFFERENTIELLE ANALYSE VON SORTING-DATEN DURCH HIERARCHISCHE CLUSTER-ANALYSEN

D i s s e r t a t i o n
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Universität Saarbrücken

vorgelegt von
HELMUT ROSSBACH
Diplom-Psychologe

Saarbrücken 1975

| | |
|---|---|
| Der Dekan: | Prof. Dr. F. Hiller |
| Die Berichterstatter: | Prof. Dr. W.H. Tack |
| | Prof. Dr. P. Orlik |
| Tag der letzten Prüfung: | 27. Juni 1975 |

UNIV.
LEIDEN
BIBL.

"The practice of taxonomy has remained intuitive and commonly inarticulate, an art rather than a science ... In the last few years there has been an increased awareness of the problems in the aims and practices of taxonomy. In particular, there has been an interest in the development of numerical methods in taxonomy as an aid to making systematics a quantitative science, a step which comes in time to every scientific discipline."

(SOKAL und SNEATH, 1963)

"Die Welt ist einmalig. Das blosse Nachsprechen der Momente, die immer und immer wieder als dasselbe sich aufdrängen, gleicht eher einer vergeblichen und zwanghaften Litanei als dem erlösenden Wort. Klassifikation ist Bedingung von Erkenntnis, nicht die selbst, und Erkenntnis löst die Klassifikation wiederum auf."

(HORKHEIMER und ADORNO, 1944)

6.3 Der dritte Datensatz

6.3.1 Die Entstehung des dritten Datensatzes

6.3.2 Die Analyse des dritten Datensatzes

6.4 Zusammenfassung der empirischen Ergebnisse

7 Diskussion und Zusammenfassung

Literaturverzeichnis

Anhang

Inhaltsverzeichnis


0 Vorwort

1 Einleitung

2 Clusteranalysen

2.1 Allgemeine historische und methodologische Vorüberlegungen

2.2 Die Funktion der automatischen Klassifikation

2.3 Das Kriterium

2.4 Der Clusterfindungsprozess

2.5 Definition des HCS

3 Die zur Klassifikation verwendeten Methoden

3.1 Die Minimum-Methode

3.2 Die Maximum-Methode

3.3 Die Methode "Gruppendurchschnitt"

3.4 Die Schwerpunkt-Methode

3.5 WARDs Methode

4 Die Logik der Kombination von Sortiermethode und HCS

4.1 Die Sortiermethode

4.2 Das Problem der inhomogenen Versuchspersonengruppe

4.3 Die Distanz von Aufteilungen

5 Der Vergleich der vier Distanz-Masse für Aufteilungen

5.1 Die empirische Erprobung der Distanzmasse

5.2 Der Versuchsplan

5.3 Erster Test

5.4 Die Diskrepanz zwischen einem HCS und einer Aufteilung (nach HUBERT, 1972)

5.5 Zweiter Test

6 Die Ergebnisse des empirischen Vergleiches

6.1 Der erste Datensatz

6.1.1 Die Entstehung des ersten Datensatzes

6.1.2 Die Analyse des ersten Datensatzes

6.1.3 Die Bewertung der Baumgraphen des ersten Datensatzes

6.2 Der zweite Datensatz

6.2.1 Die Entstehung des zweiten Datensatzes

6.2.2 Die Analyse des zweiten Datensatzes

Diese Arbeit wäre nicht ohne das wohlwollende Interesse und die Geduld Professor Werner H. Tacks entstanden und zu Ende geführt worden. Ihm ist auch für die Bereitstellung der nötigen Mittel zu danken. Die Informatiker Herr R. Zumkeller und Herr Th. Witte haben mit erfreulichem Elan alle anfallenden Berechnungen programmiert und ausgeführt. Ihnen danke ich ebenfalls an dieser Stelle.

## 1. EINLEITUNG

Unbekanntes lässt sich am besten mit Bezug auf Vertrautes erläutern. Besser als durch eine abstrakt-formale Einführung kann Sinn und Zweck clusteranalytischer Verfahren am konkreten Beispiel klargemacht werden. Wir benutzen im folgenden Gemeinsamkeiten zwischen den dem Sozialwissenschaftler vertrauten Typologien und Clusteranalysen.

Ansätze in Richtung auf eine wissenschaftlich-psychologische Typologie finden wir im Alltag, wenn wir jemanden durch eine prägnante Kennzeichnung, indem wir ihn etwa als "jünglingshaft" bezeichnen, umfassend charakterisieren. "Typologisch" nennen wir solche Wörter der Umgangssprache, weil sie jemanden zugleich in mehreren Aspekten beschreiben. So denken wir beim Typ des "Jünglinghaften" an jemanden, der überdurchschnittlich begeisterungsfähig ist und impulsiv, dabei ein wenig naiv und etwas schüchtern; wir erwarten, dass er ethisch empfindsam ist und nicht abgebrüht und kalt, usf..

> Der typologische Begriff stellt also eine ökonomische Zusammenfassung von Informationen dar.

Die begrifflichen Schubfächer, in denen Psychiater ihre Patienten versuchen unterzubringen, sind im Grund von genau derselben Art. Mit der Bezeichnung "Schizophrenie" wird etwa ein Syndrom verbunden, bestehend aus 'Nichtbeachtung der Umgebung', 'normalerweise ohne Emotionen', aber 'gelegentlich Stimmungsausbrüche ohne offensichtlichen Bezug zu Ereignissen','Halluzination' usw..

Wissenschaftliche Typologien haben häufig noch einen weitergehenden Anspruch. Während die oben angeführten Begriffe aufgrund ihres komplexen Gehalts lediglich lange Aufzählungen und Beschreibungen ersparen, streben Typologien wie die des Psychia-

ters KRETSCHMER so etwas wie ein System von Typenbezeichnungen an, eben eine Typologie im eigentlichen Sinn. Wir haben es nicht mehr nur mit einzelnen Bezeichnungen zu tun, deren Zahl unbestimmt und deren Beziehung untereinander ungeklärt bleibt, und die zudem nur der Verständigung über einzelne klinisch auffällige Personen dienen. Die vollständige Typologie stellt deutlich heraus:

1.) welche Merkmale (in diesem Fall: Körperbau und Temperament) sie zur Charakterisierung sämtlicher Typen verwendet,

2.) welcher Art die Zusammenhänge zwischen den Dimensionen sind, und versucht,

3.) alle Personen im System der festgelegten Zahl von Typen einordnen zu können.

In seiner Konzeption entsprach das KRETSCHMERsche System weitgehend diesem allgemeinen Ideal der Typologie: die ganze physische und psychische Formenmannigfaltigkeit sämtlicher Menschen sollte auf nicht mehr als vier Typen reduziert werden. Die schwierige Diagnose diverser psychischer Eigenschaften könnte so aufgrund der engen Vergesellschaftung mit Körperbaucharakteristika indirekt erfolgen, nämlich über den Schluss, dass das Vorhandensein des einen Merkmals das Vorhandensein des anderen impliziere oder doch sehr wahrscheinlich mache. Es ist kein Geheimnis, dass diese und andere klassische Typologien in den Sozialwissenschaften nicht gehalten haben, was man sich von ihnen versprach. Das kam zunächst darin zum Ausdruck, dass die Einordnung der Individuen Schwierigkeiten bereitete, sei es, dass nur ein kleiner Anteil der Gesamtpopulation wirklich hinreichend "typisch" war, um eine sichere Einordnung zu gestatten, (die überwiegende Mehrheit fiel also in einen ungenügend definierten Zwischenbereich) oder dass dieselbe Person - wie das in der Psychiatrie noch heute der Fall ist - von verschiedenen Diagnostikern verschieden kategorisiert wurde, weil scharfe Grenzen zwischen den Typen zu ziehen unmöglich war.

Wo liegen nun die Gründe für das Versagen der klassischen Typologie ?

Das lässt sich anhand der eingangs erarbeiteten drei Kriterien leicht erläutern.

Zu 1.) Wir definieren den Typus durch ein charakteristisches Bündel von Merkmalen. Es dürfte einleuchten, dass der komplexe Begriff des Typus nur soweit brauchbar ist, als seine Konstituenten selbst einfacherer Natur, das heisst aber im wesentlichen leichter messbar sind. Dies war offensichtlich bei den klassischen Typologien nicht gegeben; weder die entsprechenden Körperbaumerkmale und noch viel weniger die psychologischen Dimensionen waren exakter erfassbar als die Typen selbst, die intuitiv konzipiert wurden. Notwendig blieben daher die Typusbegriffe so vage wie die zugrundeliegenden Merkmale.

Zu 2.) Selbst präzis erfasste Merkmale wären nicht ausreichend zur Etablierung einer Typologie, wenn nicht die Messdaten der Population eine besondere Eigenschaft aufweisen. Damit viele Individuen wirklich zureichend durch wenige Kategorien (eben die Typen) beschreibbar sind, müssen charakteristische Muster von Merkmalsausprägungen überzufällig häufig sein. Wenn in der zugrundeliegenden Population die Werte auf den einzelnen Dimensionen völlig unabhängig voneinander variieren, ist die Aufzählung sämtlicher Messwerte für jede Person nicht durch die ökonomischere Charakterisierung durch Zuweisung zu einem aus einer geringeren Anzahl von Typen zu ersetzen.

Zu 3.) Das Ziel, möglichst alle Personen einer Population einem der Typen eindeutig zuordnen zu können, verlangt eine deutlich diskontinuierliche Strukturierung.

Die Abkehr von der kategorialen und die Zuwendung zur Quantitativen Beschreibung in den Sozialwissenschaften erschien daher nur folgerichtig. Aber nachdem, dank einer entwickelten quantitativen Methodik, über längere Zeit erfolgreich grosse Mengen an Messdaten angehäuft worden waren, entstand wiederum ein Bedürfnis nach datenreduzierenden Verfahren, die durch Strukturierung die Informationsfülle überschaubarer und besser handhabbar gestalten sollten. Diese Funktion wird heute praktisch ausschliesslich durch die Faktorenanalyse erfüllt.

Wenn inzwischen wieder ein steigendes Interesse an typologischen Techniken zu verzeichnen ist, so hat das sicher nicht nur den sozusagen "subjektiven" Grund, dass unsere Sprache nun einmal digital und nicht "analog" kodiert, und deshalb das Hantieren mit quantitativen Grössen (was ja die Faktorenwerte auch wieder sind) immer etwas umständlich ist. Es kann vielmehr auch objektive Gründe geben, eine Clusteranalyse vorzuziehen, Gründe, die in der Struktur der Daten selbst liegen. Ein Beispiel aus der Geologie mag verdeutlichen, welche Art von Strukturiertheit gemeint ist. Bodenkundler sind in der Lage, nach der Untersuchung von Bodenproben exakte Messwerte für verschiedene Eigenschaften anzugeben. Obwohl man sich also auf differenzierte Angaben über Wasser- und Humusgehalt, Raumgewicht, Säuregrad, Anteil verschiedener Substanzen wie etwa Kalk, usw. stützen kann, wird man trotzdem, jedenfalls für gewisse Zwecke, statt in jedem Fall alle Messwerte aufzulisten, sich durch Verwendung von Bezeichnungen für typische Bodenarten (Heideboden, Lössboden usw.) verständigen. Man tut dies im Bewusstsein, dass die Ausprägungen auf diesen Dimensionen nicht in allen denkbaren Kombinationen auftreten, sondern dass einige völlig ausgeschlossen sind, andere dafür überzufällig häufig vorkommen.

Wieso können wir damit rechnen, dass trotz kontinuierlicher Messgrössen nicht alle numerisch und logisch möglichen Kombinationen auftreten ? Für unser geologisches Beispiel ist das leicht verständlich zu machen:

Es mögen etwa bestimmte Säuregrade mit bestimmten Anteilen von Kalk und anderen Mineralien unverträglich sein, oder ein bestimmter Humusanteil impliziert gewisse physikalische Eigenschaften. Eher aus subjektiven Gründen erklärt sich die Tatsache, dass Mediziner sich praktisch ausschliesslich mit Hilfe kategorialer Krankheitsbezeichnungen verständigen. Die Fülle präziser Messwerte über physiologisch-chemische Grössen, die ja den Einzelfall viel genauer beschreiben, werden allenfalls zur Einordnung in eine vorgegebene Kategorie benutzt oder bei "Grenzfällen" extra angeführt. Die sind aber anscheinend so selten, dass das Klassifikationsschema dadurch seine Brauchbarkeit nicht verliert.

Im sozialwissenschaftlichen Bereich sind ganz ähnliche funktionale Abhängigkeiten zwischen verschiedenen Variablen zu erwarten: Beim Bemühen um eine Industrialisierung sogenannter unterentwickelter Regionen zeigt sich, dass ein ganzes Bündel von bestimmten Einstellungen, Gewohnheiten und ähnliches gegeben sein müssen, damit sich irgendwelche "Institutionen" verpflanzen lassen. Offenbar bilden sich unter dem Einfluss von Selektion und sozialen Lernprozessen spezifisch optimale Kombinationen verschiedener Grössen in einer Art Verflochtenheit, die die isolierte Betrachtung oder auch Beeinflussung e i n e s Merkmals unsinnig erscheinen lässt.

Veranschaulicht man sich diesen Sachverhalt vereinfacht geometrisch, ergibt sich etwa dies Bild für zwei Dimensionen:

Abb. 1: Punkteverteilung im zweidimensionalen Messwertraum, die eine Unterscheidung von drei Teilmengen (empirische Typen) nahelegt. Jede dieser Gruppen von einander ähnlichen Objekten ist deutlich von den übrigen separiert.

Die zwar logisch denkbaren aber funktional nicht möglichen Fälle machen den leeren Raum um verschiedene Dichtezentren aus. Wir finden Punktwolken unterschiedlicher Grösse und Ausgefülltheit, die als Typen anzusprechen wären. Im Unterschied zu den Typen der klassischen Typologien (Ideal- und Extremtypen) sollen sie als empirische Typen bezeichnet werden. Empirische Typen sind definiert als abgrenzbare Bereiche im Messwertraum, in denen sich Objekte überzufällig häufen. Ihnen entsprechen Merkmalskombinationen auf einer u.U. grösseren Zahl von Dimensionen, die aus irgendwelchen inhaltlichen Gründen eher in der Wirklichkeit auftreten als andere mathematisch ebenso gut denkbare.

Das empirische Typenkonzept und damit auch die Verfahren der Typenfindung, im weiteren Clusteranalysen genannt, stehen also nicht in Gegensatz zu Skalierungs- oder Dimensionsanalytischen Verfahren. Sie ergänzen sie, beziehungsweise bauen auf ihnen auf.

## 2. CLUSTERANALYSEN

### 2.1 Allgemeine historische und methodologische Vorüberlegungen

Die clusteranalytischen Modelle sind der Versuch einer Formalisierung von Vorgängen, wie sie etwa auch aller Ding- und Eigenschaftswahrnehmung (cognition) und damit ganz wesentlich der verbalen Kommunikation zugrundeliegen, nämlich der 'Konstitution eines Gegenstands als Gegenstand' oder schlichter gesagt der Klassenbildung. Daher treten in diesem, auch "automatische Klassifikation" genannten, Bereich auch die von Philosophen, Psychologen und Physiologen unter so verschiedenen Themenbezeichnungen wie "Hermeneutik", "Induktion", "concept formation" und "Invariantenbildung" abgehandelten Probleme auf.

Ein historisch bedeutsamer Zweig typologischen Denkens leitet sich her von Platons Ideenlehre, nach der die beobachtbaren Gegenstände unvollkommene und wechselnde Abbilder unveränderlicher ewiger Ideen sind. In dieser "idealistischen" Tradition stehend, wird Goethes Typenlehre gesehen, wie er sie in dem "Ersten Entwurf zur allgemeinen Einleitung in die vergleichende Anatomie (1796)" entwickelt. Dort empfiehlt er den Vergleich der empirischen Formen mit dem "idealen Typus" als Forschungsstrategie. Seine Vorstellungen von diesem Idealtypus sind allerdings vage und mehrdeutig, wie die beiden folgenden Zitate (nach Remané, 1956) zeigen:

"Hierbei fühlte ich bald die Notwendigkeit, einen Typus aufzustellen, an welchem alle Säugetiere nach Übereinstimmung und Verschiedenheit zu prüfen wären, und, wie ich früher die Urpflanze aufgesucht, so trachtete ich nunmehr, das Urtier zu finden, denn doch zuletzt: den Begriff, die Idee des Tieres!"

"Wie nun aber ein solcher Typus aufzufinden, zeigt uns der Begriff desselben schon selbst an: Die Erfahrung muss uns die Teile lehren, die allen Tieren gemein und worin diese Teile bei verschiedenen Tieren verschieden sind, alsdann tritt die Abstraktion ein, sie zu ordnen und ein allgemeines Bild aufzustellen."
Goethe, 1796

Während im letzten Zitat schon eine moderne empirische Auffassung anklingt, kehrt ein später Nachfahre dieser idealistischen Tradition wieder zum erfahrungswissenschaftlich informationsleeren platonischen Typusbegriff zurück. NAEF, der noch 1919 ein Buch mit dem Titel "Idealistische Morphologie und Phylogenetik" veröffentlichte, schrieb 1931:

> "Wir nennen diejenigen Naturformen, die sich auf einen gemeinsamen T y p u s beziehen lassen, "formverwandt" oder "typisch ähnlich". Der Begriff der "Formverwandtschaft" oder "typischen Ähnlichkeit" bezeichnet ein ideelles, rein formales Verhältnis beobachteter Bildungen innerhalb einer natürlichen Gruppe, .... Andererseits kann die natürliche systematische Kategorie definiert werden als die Gesamtheit der auf einen bestimmten Typus zurückzuführenden Formen."

Am Beispiel dieses Zitates erweist sich die Zirkularität der naturwissenschaftlichen Anwendung der platonischen Idee: Die typologische Ähnlichkeit wird definiert auf der Basis schon bestehender "natürlicher Gruppen", ohne dass gesagt wird, warum gerade diese Gruppierung der Objekte natürlich sei oder wie sie zustande gekommen ist.

Doch gerade Goethe wird man mit einer so pauschalen Verurteilung nicht gerecht. Konrad LORENZ weist in seiner Abhandlung über "Gestaltwahrnehmung als Quelle wissenschaftlicher Erkenntnis (1959)" und später in "Die Rückseite des Spiegels (1973)" darauf hin, dass die spontane Invariantenbildung im wesentlichen ohne jede bewusste analytische An-

strengung zahlreiche Einzeldaten und komplexe Beziehungen zwischen ihnen in einer Weise zu erfassen und zu verarbeiten in der Lage ist, wie wir sie uns in der Clusteranalyse bewusst machen.

Im Zusammenhang mit den mehr psychologischen Voraussetzungen von Wissenschaft, aber auch bei formalen Untersuchungen zum Induktionsproblem, zeigt sich, dass GOETHE etwas beschreibt, was keineswegs durch bessere Einsichten überholt worden ist, auch nicht durch die moderne, heute wohl schon etwas relativierte, Bevorzugung der Quantifizierung.

In Ergänzung zu LORENZ' eher phänomenologischer Betrachtungsweise steht WATANABEs (1969) formallogische Darstellung der jeder Induktion und jeder Quantifikation notwendig vorausgehenden Konzeptbildung.

> Eine Hypothese der Form "alle X sind P" (wie z.B. "alle Raben sind schwarz") beginnt mit dem kognitiven Akt der Anerkennung der Existenz einer Gruppe ähnlicher Objekte, denen der Name X gegeben wird. Angenommen (m + n) Eigenschaften werden von allen vorliegenden Mitgliedern von X geteilt, und m Eigenschaften werden als ausreichend angesehen zur Identifikation eines Objektes als Mitglied der Klasse X, dann können wir die m Eigenschaften zur Charakterisierung von X nehmen und irgendeines der übrigen aus n als P und die Hypothese aufstellen "alle X sind P". Dieser Prozess erfordert die Auswahl von Eigenschaften auf zwei verschiedenen Stufen. Als erstes verlangt die Erkennung der Gruppe schon eine besondere Auswahl wichtiger Eigenschaften .... Auf der zweiten Stufe müssen dann die Kerneigenschaften und ein bestimmtes P aus den übrigen n Eigenschaften .... gewählt werden. Tatsächlich ist für jede endliche Menge von Objekten die Zahl der Eigenschaften, die logisch für X und P möglich sind, ungeheuer gross, aber nur sehr wenige von ihnen können ernstlich für eine Induktion in Betracht gezogen werden. " (Übersetzt vom Verfasser)

Was hier über die Klasse X gesagt wurde, gilt für jeden empirischen Begriff. Von den (m + n) Eigenschaften - und möglichst noch einigen mehr - e r w a r t e n wir, dass sie realiter miteinander verbunden sind, obwohl es dafür einen hinreichenden logischen Grund nicht gibt. Wäre das nämlich der Fall, so hätten wir es lediglich mit einer Nominaldefinition, d.h. einer Tautologie ohne jeden Informationsgehalt zu tun.

Sinnvoller noch als die dem zugrundeliegende scharfe Dichotomie von logischer Notwendigkeit und faktischer Gegebenheit (analytischer und synthetischer Wahrheit) ist jedoch POPPERs (1966) komparativische Darstellung ihrer Beziehung als eine wechselseitige, reziproke Abhängigkeit von "logischer Wahrscheinlichkeit" und "empirischem Gehalt".

Danach ist ein Begriff umso informativer und fruchtbarer, je mehr nur empirisch einzulösende Zusammenhangsbehauptungen er enthält beziehungsweise nahelegt - 'projectability' im Sinne GOODMANs (1955). Erfahrungswissenschaftlich-deskriptive Begriffe sind also Klassifikationen,die als ein Bündel spezifischer Hypothesen über eine bestimmte "Ordnung" in der Natur aufgefasst werden können. Erweisen sie sich als fruchtbar, so ersparen sie uns redundantere Beschreibungen unserer Gegenstände, indem sie die "zu erwartende Anzahl von Unterscheidungen, die zur Charakterisierung eines zufällig ausgewählten Objektes nötig sind, minimieren" (nach HAWKINS, 1968, S. 44, übers. v. Verf.). Damit eine solch ökonomische Beschreibung vom Gegenstand her überhaupt möglich sein kann, muss natürlich diese Ordnung faktisch existieren, d.h., dass statt aller logisch möglichen tatsächlich nur die mit dem Klassifikationssystem hypostasierten Merkmalskombinationen auftreten. Der G r a d des empirischen Gehalts bemisst sich konkret am numerischen Verhältnis dieser beiden Mengen zueinander.

Eine zunächst abstrus anmutende Zuspitzung erfahren diese Überlegungen in WATANABEs "Theorem vom hässlichen Entlein", das so etwas wie die Anwendung des GOODMANNschen Paradoxons der Induktion auf das Klassifikationsproblem darstellt. (N. GOODMANN hatte gezeigt, dass dieselben Daten Hypothesen mit einander widersprechenden Konsequenzen induktiv in gleicher Masse stützen können. Wenn diese verwirrende Tatsache normalerweise nicht zum Problem wird, so liegt das an der inhaltlich begründeten Bestimmung dessen, was induktionsfähige Prädikate sind. Rein logisch gibt es jedoch keine zureichenden Kriterien für deren Auswahl und damit auch keine Möglichkeit, diesen absurden Folgen zu entgehen.) In Gegensatz zur Auffassung der Klassiker der Taxonomie (SOKAL und SNEATH, 1963) stellt sich WATANABE mit seinem Beweis, dass die Bedingung der Möglichkeit jedweder Klassifikation die ungleiche Gewichtung der die zu klassifizierenden Objekte charakterisierenden Merkmale ist, und dass es für diese Gewichtung keine formallogische Begründung gibt. Nach diesem Theorem ist die Welt zunächst homogen grau, weil alle Objekte untereinander gleich ähnlich sind, d.h. jedes Objekt mit jedem anderen die gleiche Anzahl von Merkmalen * teilt und sich von jedem in der gleichen Anzahl unterscheidet. Dies gilt für jede Menge (nicht identischer) Objekte und vor allem für jede, wie auch immer ausgewählte Merkmalsmenge, sofern sie einen Booleschen Verband bildet. Tut sie das nicht, so argumentiert WATANABE, dann sei bereits eine Gewichtung vorgenommen. Denn unter logischem Gesichtspunkt gibt es keine "natürlichen", "primären" oder "atomaren" Eigenschaften, weil verschiedene Teilmengen des Booleschen Verbandes in gleicher Weise die Funktion, die übrigen Merkmale zu generieren, übernehmen können. Das schliesst - wenigstens aus formalen Gründen - die Bevorzugung einer bestimmten als der "grundlegenden" Teilmenge aus.

Welche Bedeutung kommt nun einer so offensichtlich weltfernen These zu ? Denn tatsächlich gibt es für uns natürlich

---

* WATANABE demonstriert und beweist sein Theorem nur für binäre Merkmale, behauptet aber, dass es sinngemäss auch für kontinuierliche Variablen gilt.

unterschiedlich ähnliche Objekte. Wir verwechseln leicht einmal Ente und Gans, haben aber keine Schwierigkeiten, diese vom Tannenbaum zu unterscheiden. Im Licht von WATANABEs Theorem stellt sich dieses unser Vermögen so dar: Immer dann, wenn wir zwei Objekte als ähnlicher als ein anderes Paar bestimmen, haben wir entweder den von ersteren geteilten Merkmalen eine grössere Bedeutung beigemessen als denjenigen, die letzteren gemeinsam sind - und ihre Anzahl ist jedesmal dieselbe - oder diese Anzahl ist in beiden Fällen verschieden, weil nur eine Teilmenge des vollständigen Booleschen Verbandes zu ihrer Charakterisierung herangezogen wurde. In beiden Fällen handelt es sich um eine formal nicht begründbare Gewichtung.

Welches ist nun aber die Basis für die faktisch geleisteten Ähnlichkeitswahrnehmungen und -urteile bzw. für die ihnen zugrunde liegende Auswahl und Gewichtung der Prädikate ? Nach GOODMAN so etwas wie deren alltägliche Bewährung, nach WATANABE unser theoretisches Wissen über den betreffenden Gegenstandsbereich und nach Konrad LORENZ der in der Phylogenese akkumulierte Erfolg mit derartigen Differenzierungen.

Es liegt auf der Hand, dass hier keine echten Auffassungsgegensätze bestehen. Die Autoren legen die Betonung, so könnte man ihre Standpunkte integrierend sagen, einfach auf verschiedene Quellen unseres Vorwissens. Dieses Vorwissen kann so einfach und selbstverständlich sein wie die stammesgeschichtlich im Organismus verankerte Fähigkeit bestimmte Farben zu unterscheiden oder so komplex wie das Vermögen, die Idee eines "Bauplanes" zu entwickeln, der hinter den, wie immer im einzelnen defekten, Realisationen "gesehen" wird.

Fazit: "Automatische Klassifikation, als so objektiv - neutral sie oft gedacht wird, ist weder methodisch noch inhaltlich voraussetzungslos. Tatsächlich fügt sie sich sehr gut dem Bild vom "hermeneutischen Zirkel": Von unserem Vorwissen geleitet wählen wir Objekte und Merkmale aus. Auch den Ergebnissen des Klassifikationsprozesses trauen wir nicht blind, sondern messen sie an unseren Erwartungen, um beide dann in wechselseitiger Korrektur zu verbessern.

## 2.2 Die Funktion der automatischen Klassifikation

Die Klassifikation eines Objektbereichs, sei es die uns mit den Begriffen der Umgangssprache gegebene oder das Ergebnis eines taxonomischen Verfahrens, hat einen Zweck zu erfüllen und ist daran zu messen. Das Kriterium, das die Wissenschaftstheorie für terminologische Systeme entwickelt hat, ist deren Fruchtbarkeit.

Wir wollen auf die Problematik dieses Konzepts nicht weiter eingehen als es die Absicht, eine gewisse Paralellität zu den Kriterien clusteranalytischer Modelle aufzuweisen, erforderlich macht. Wenn wir im Alltag von jemandem sagen, er sei ein "unbelehrbarer Nazi", so benutzen wir zu seiner Charakterisierung eine Typenbezeichnung aus einer Typologie politischer Orientierung und verknüpfen auf diese Weise unter Umständen recht vielfältige Erwartungen mit dieser Person. Die Fruchtbarkeit der Bezeichnung bemisst sich nach dem Umfang und der Zuverlässigkeit dieser Rückschlüsse. Man hat, wenn man derart in einem Begriff gebündelte Informationen weitergibt, nichtlineare Prädiktion vorgenommen - und das ist auch das Vorgehen der Clusteranalysen: Es werden die Objekte eines bestimmten Erfahrungsbereichs entsprechend ihrer Ähnlichkeit unter eine kleinere Zahl von Oberbegriffen subsumiert, die dann eine Typologie bilden, so wie sich in unserem Beispiel jeweils unbelehrbare Nazis, Liberale, Linksradikale usf. bezüglich verschiedener in diesem Zusammenhang bedeutsamer Eigenschaften wie Vorstellungen über die richtige Staatsordnung, angemessene Mittel in der Politik, Erziehungsstil, Verhältnis zum anderen Geschlecht usw. ähneln sollen.

Bevor wir entsprechende Überlegungen auf taxonomische Verfahren beziehen können, müssen wir im einzelnen erörtern, von welcher Art Einengung und Präzisierung sind, die hier wie in jedem anderen Fall mit mathematischer Formalisierung verbunden sind.

Hervorzuheben ist als eine wesentliche Beschränkung aller numerischen Klassifikationsmethoden vor allem die folgende

Eigentümlichkeit:

Von umgangssprachlichen Begriffen unterscheiden sich empirische Typen durch das Fehlen von "Offenheit". Im Vergleich zu den Begriffen der Umgangssprache ist den wissenschaftlichen Termen die Kontextabhängigkeit der Bedeutung weitgehend durch intensionale Definition genommen. Sie bleiben aber offen in dem Sinn, dass die sie definierende Merkmalskombination prinzipiell die Neuzuordnung weiterer Objekte gestattet. Im Gegensatz dazu sind die methodisch gewonnenen Klassen stets endliche Mengen, und der so bestimmte Typus oder Begriff - wenn man von einem solchen überhaupt noch reden will - ist ein in diesem eingeschränkten Sinn extensional definierter.

Diese Beschränkung ergibt sich als logische Konsequenz aus dem Umstand, dass Basis für den taxonomischen Prozess immer endliche Mengen von Objekten sind, mögen diese Elemente auch, wie vor allem in biologischen Anwendungen, als 'Exemplare einer Art' gelten oder wie in psychologischen Zusammenhängen selbst offene Begriffe sein und dadurch in der anschliessenden Interpretation die Offenheit von Klassenbezeichnungen begründen.

Eine erste Differenzierung der Methoden orientiert sich an der Form der Klassen, die bei ihrer Anwendung entstehen. Um mit dem strengsten und zugleich wichtigsten Fall zu beginnen:

1. Die Klassen sind disjunkt (d.h. die Mitgliedschaft eines Objektes in der einen, schliesst die in irgendeiner anderen aus), und zugleich vollständig (d.h. jedes Objekt gehört einer Klasse an. Unter Umständen besteht diese auch nur aus ihm selbst).

Zwei weitere Formen von Klassifikation ergeben sich, wenn man zunächst die Bedingung der Vollständigkeit und dann auch die der Disjunktivität aufhebt.

2. Die Klassen sind disjunkt, aber nicht notwendig vollständig. Ein Teil der Objekte wird also gar keiner Klasse zugewiesen.

3. Es gibt überhaupt keine Restriktionen bezüglich der Art der Klassen. Die Klassen können also auch "überlappen", d.h. ihre Schnittmengen müssen nicht leer sein.

Der denkbare 4. Fall, dass wohl Vollständigkeit, nicht jedoch Disjunktivität gegeben ist, hat unseres Wissens keine Bedeutung gewonnen.

Klassifikationen vom Typ 1 nennen wir "Aufteilungen"(engl. "Partitions"). Die Aufteilung P einer Menge S ist also definiert als eine Menge von Teilmengen $c_1, c_i, c_j, \ldots c_t$ mit

Definition 1:

| | | |
|---|---|---|
| A1 | $c_i \neq \emptyset$ | |
| A2 | $i \neq j \Rightarrow c_i \cap c_j = \emptyset$ | Disjunktivität |
| A3 | $\bigcup_{i=1}^{t} c_i = S$ | Vollständigkeit |

Es lassen sich nun zwei extreme Aufteilungen vorstellen:

1.) Jene Aufteilungen, die alle Elemente einer einzigen Klasse zuweist, bei der also $t = 1$ ist,

2.) eine Aufteilung, die nur einelementige Mengen als Klassen enthält, bei der mithin $t = |S|$ ist.

Aus leicht ersichtlichen Gründen sind sie trivial, aber aus grundsätzlichen Überlegungen nicht uninteressant. Die erste ist nämlich gleichbedeutend mit der Unmöglichkeit, irgendwelche Unterschiede zwischen den Objekten zu konstatieren. Aber irgendeine Art von Unterscheidung, und sei es die in zu einem Typ gehörige Objekte und nicht dazugehörige, also mindestens eine Zweiteilung macht erst eine sinnvolle Klassifikation aus und bildet die Voraussetzung für informative Aussagen. Das andere Extrem tritt ein, wenn individuelle Differenzen so wichtig genommen werden, dass letztlich nur Eigennamen vergeben werden, aber die zu Oberbegriffen oder Klassenbezeichnungen führende Abstraktion nicht vollzogen wird. Die für jede sinnvolle Verständigung und erst recht die nomothetische Wissenschaft charakteristische Leistung, nämlich von Individuenkonstanten zu Allgemeinbegriffen aufzusteigen, wird damit nicht erbracht.

Dazwischen liegt nun eine unüberschaubare Anzahl von Aufteilungen, unter denen sich die eigentlich interessanten Fälle befinden. Welche dies sind, entscheidet sich nach Massgabe eines Kriteriums. Die diesem Kriterium genügenden Aufteilungen sollen aber nicht nur bestimmt sondern auch gefunden , d.h. namhaft gemacht werden. Diese beiden Funktionen in sich zu vereinen, macht ein Klassifikationsverfahren aus.

## 2.3 Das Kriterium

Statt uns gleich auf diverse mathematisch formulierte Kriterien einzulassen, bemühen wir zunächst lieber unser Vorverständnis darüber, was eine Aufteilung als brauchbar ausweisen könnte.

Welche Gründe veranlassen uns, verschiedene Gegenstände mit demselben Begriff zu belegen ? Offenbar ist es deren - jedenfalls in einem bestimmten Rahmen gegebene - Austauschbarkeit. Dem Begriff "Austauschbarkeit" haftet ein hier durchaus erwünschter pragmatischer Beiklang an: Wenn wir im Alltag zum Beispiel so verschiedenes wie Töne, Lichtaufleuchten und Armbewegungen als "Signal" bezeichnen, dann, weil sie für einen gegebenen Zweck austauschbar sind. In einem anderen pragmatischen Zusammenhang würde, was in einem Fall durch dieselbe Bezeichnung zusammengefasst war, als etwas Verschiedenes klassifiziert. So ist es nicht nur anatomische Unbeschlagenheit, die den Seemann vom Walfisch reden lässt, sondern sein vom systematisierenden Zoologen gänzlich verschiedenes Interesse. Beide, Seemann und Zoologe, sind bei ihrer Klassifikation durch Austauschbarkeit geleitet. Für den einen sind die möglichen Verrichtungen nicht sehr verschieden, ob er es nun mit einem Hai oder einem Grindwal zu tun hat, der andere hat im mehr theoretischen Sinn Austauschbarkeit im Auge.

Wie können wir unser intuitives Konzept "Austauschbarkeit" für taxonomische Zwecke präzisieren ? Da wir nicht die Absicht haben, einen vollständigen Überblick zu geben, sondern nur Grundlagen zum Verständnis des Folgenden bereiten wollen, begnügen wir uns mit Grundsätzlichem und Beispielhaftem.

Allgemein gesprochen ist eine Aufteilung dann befriedigender

als eine andere, wenn sie die Objekte so gruppiert, dass Vorhersagen über sie aufgrund der Kenntnis der Klassenzugehörigkeit weniger fehlerhaft sind.

Diese Bestimmung berücksichtigt noch nicht, in welcher Form uns Informationen über die Objekte gegeben sind. Prinzipiell gibt es da zwei Möglichkeiten:

1.) als Objekt/Prädikattafel.
In diesem Fall liegt für jedes Objekt eine Angabe über deren Wert auf der für alle gleichen Menge von Merkmalen vor.

2.) als reellwertige Funktion auf der Menge der Objektpaare.
In einer Matrix findet sich ein Wert für jedes Objektpaar als für deren Ähnlichkeit bzw. Distanz. *

Bei dem Versuch unseren Definitionsansatz auf den ersten Fall anzuwenden, erweist es sich als notwendig, bezüglich der Skalenqualität in der die Eigenschaften der Objekte gemessen wurden, zu differenzieren, weil die Möglichkeiten, den Begriff des Fehlers genauer zu bestimmen, davon abhängen. Um nur einige dieser Möglichkeiten zu skizzieren: Bei kategorialen Daten (Merkmal vorhanden/nicht vorhanden) liegt es nahe, die Aufteilung nach dem Grad der Ordnung zu bewerten, den sie im informationstheoretischen Sinne unter den Objekten herstellt. Mit grösserer Ordnung sinkt die Wahrscheinlichkeit einer falschen Vorhersage über das Vorhandensein einer bestimmten Eigenschaft, weil die Wahrscheinlichkeit dafür, dass ein Objekt diese besitzt, nach Kenntnis seiner Klasse entweder sehr klein oder sehr gross ist. Wenn die Daten - um gleich zum anderen Extrem überzugehen - als mindestens intervallskalierte Messwerte vorliegen, kann es sinnvoll sein, als Fehler den Betrag der Differenz zwischen dem Messwert des Objektes und

---

* Für die folgenden Überlegungen ist es unerheblich, dass die meisten Verfahren auf Daten der Form 2) operieren und dass es diverse Möglichkeiten der Transformation von 1) nach 2) gibt.

der durch unser Wissen um seine Klassenzugehörigkeit vermittelten Vorhersage anzusehen. Die beste Schätzung ist der Mittelwert der betreffenden Klasse, und die beste Aufteilung demnach die mit minimaler Binnenvarianz über alle Merkmalsdimensionen.

Die Funktion obiger Kriterien ist es, schlicht gesagt, Aufteilungen mit homogenen Klassen zu favorisieren und homogen sind sie, wenn ihre Elemente untereinander sehr ähnlich sind. Mit dieser Zusammenfassung ist zugleich die Verbindung zum zweiten Fall hergestellt. Und es erhebt sich die Frage, wie wir mit Bezug auf Distanzen oder Ähnlichkeitskoeffizienten die Homogenität von Klassen und damit die Güte einer Klassifikation definieren wollen.

Man kann im Sinne des zu 1) Ausgeführten die Binnen-Streuung zum Mass nehmen. Aber es lassen sich bei entsprechender Zielsetzung bzw. Datenstruktur auch andere Operationalisierungen von Homogenität denken, wie zum Beispiel die maximale Distanz innerhalb der Klasse oder, was durchaus etwas ganz anderes bedeuten kann, die maximale Distanz zur nächsten anderen Klasse.

Man kann sich auch daran orientieren, dass eine Aufteilung ebenfalls als Distanzmatrix darstellbar ist mit

A'1 $\quad d_{ij} = 0$ für $i,j \in c_l$ und

A'2 $\quad d_{ij} = 1$ für $i \in c_k$, $j \in c_l$, $k \neq l$,

und die Klassifikation danach bewerten, wie ähnlich diese Distanzmatrix der Ausgangsmatrix ist. Dann definiert ein Ähnlichkeitsmass für Distanzmatrizen das Kriterium.

Die naheliegende Spezifikation dieses Ähnlichkeits- oder dann richtiger Distanzmasses als Summe der Differenzbeträge zwischen den einander entsprechenden Zellen aus beiden Matrizen führt wieder auf das bereits erwähnte Streuungskriterium. Dies ist auch bestimmend, wenn man im selben Sinn wie bei den

dimensions-analytischen Verfahren (Faktorenanalyse, Multidimensionale Skalierung) die "Güte" einer Aufteilung durch den Anteil der erklärten Varianz misst. Der Anteil der erklärten Varianz erhöht sich bei Aufteilungen in dem Masse wie die Binnenvarianz der Klassen sinkt, beziehungsweise ihre Zwischenvarianz zunimmt, die zusammengefassten Objekte also einander ähnlicher sind.

Mit diesen Darlegungen dürfte deutlich geworden sein, was auch schon auf der Stufe der Auswahl von Objekten und Merkmalen - also auf inhaltlicher Ebene - durch Parallelsetzung mit dem Induktionsproblem gezeigt wurde: auch aus formalen Gründen kann es d i e richtige Aufteilung nicht geben. Das Unpersönlich--Mechanische der "automatischen Klassifikation" darf nicht mit Neutralität gegenüber Zwecken verwechselt werden. Wenn verschiedene Methoden verschiedene Lösungen ergeben, so kann das an unterschiedlichen Suchstrategien, aber auch an verschiedenen Kriterien liegen. Mit der Wahl zwischen Methoden verbindet sich also meist auch die zwischen verschiedenen Kriterien.

## 2.4 Der Clusterfindungsprozess

Mit der Entscheidung für ein bestimmtes Kriterium ist aus zwei Gründen noch nicht viel erreicht.

Erstens gilt für fast alle Kriterien und Daten, dass die triviale Aufteilung, bei der jedes Objekt eine Klasse für sich bildet, die beste ist. Das wird verständlich, wenn man sich zum Beispiel am Anteil der erklärten Varianz orientiert und die Parallele zu anderen Methoden der Datenreduktion zieht:
Bei den dimensionsanalytischen Verfahren wächst die Güte der Anpassung mit der Zahl der Dimensionen, bei Aufteilungen dementsprechend mit der Anzahl der Klassen. Hier wie da liegt die Lösung in einem sinnvollen Kompromiss zwischen dem Ziel der Vereinfachung und der Genauigkeit der Abbildung von in den Ausgangsdaten bestehenden Relationen.

Es liegt nahe, von derjenigen der beiden Zielsetzungen auszugehen, die für ein gegebenes Problem den Vorrang besitzt.

Man könnte also den Grad der angestrebten Vereinfachung vorgeben, indem man die Zahl der Klassen im voraus festlegt. Mit der Wahl eines bestimmten Kriteriums ist theoretisch zwar nun auch entschieden, welche Aufteilung in k Klassen die beste ist. Aber damit ist noch nicht viel gewonnen, denn die möglichen Aufteilungen liegen ja gar nicht zum Vergleich vor. Es ist leicht einzusehen, dass es auch kein sinnvolles Unterfangen wäre, sie alle systematisch aufzulisten. Ihre immense Zahl - und das ist nun der zweite Grund - schon für kleine Mengen (n > 20) verbietet jeden Gedanken daran, auf diesem Weg eine Lösung zu suchen.

Die Anzahl der Aufteilungen von n Objekten auf m Klassen ist eine Stirling-Zahl zweiter Art (ABRAMOWITZ und STEGUN, 1968).

$$(1) \qquad \left\{ {n \atop m} \right\} = \frac{1}{m!} \sum_{k=0}^{k=m} (-1)^{m-k} \binom{m}{k} k^n .$$

Zur Veranschaulichung der Grössenordnung, die diese Zahl schnell annimmt, ein Beispiel. Bei nur 25 Objekten und 5 Klassen gibt es bereits

$$\left\{ {25 \atop 5} \right\} \approx 2{,}4 \cdot 10^{14}$$

mögliche Aufteilungen.
Für Zweiteilungen vereinfacht sich der Ausdruck (1) zu

$$(2) \qquad \left\{ {n \atop 2} \right\} = \frac{1}{2} \quad (2^n - 2).$$

Eine für Clusterverfahren nicht unrealistische Aufgabe wäre die optimale Zweiteilung einer hundertelementigen Menge. Das ergäbe nach (2) ca. $6{,}3 \cdot 10^{29}$ Möglichkeiten. Nach einer Schätzung von WALLACE (1968) brauchte der schnellste existierende Computer für ihren Vergleich $10^{17}$ Jahre !

Der so rapide mit der Anzahl der Objekte anwachsende Umfang der nötigen Rechenoperationen bildet gleichermassen die Grundlage für die Kritik einer gegenläufigen Strategie. Sie bestünde darin, eine obere Grenze für den Grad des akzeptierten Informationsverlustes festzulegen, ausgedrückt als Wert eines Kriteriums wie zum Beispiel 'Anteil der erklärten Varianz'. Man könnte sich vorstellen, dann über die Durchsicht aller Zwei-, Drei-, Vierteilungen usw. schliesslich zu einer Aufteilung zu gelangen, die erstmals dem Kriteriumswert genügt.

Beide Wege basieren auf vollständiger Auflistung (sei es, dass sie von fester Klassenzahl m ausgehen und das Kriterium optimieren oder dass sie umgekehrt bei sukzessiver Vergrösserung von m sich einem vorbestimmten Kriterium nähern) und sind deshalb praktisch zum Scheitern verurteilt. Daraus resultiert nun der vierte Grund für die Ambiguität der numerische Klassifikation (der dritte ist die am Anfang dieses Abschnitts erläuterte Kompromisshaftigkeit, die notwendig jede Entscheidung für eine konkrete Aufteilung kennzeichnet): Die Suche nach der akzeptablen Aufteilung kann nur eine heuristische sein. Da es aus numerischen Gründen unmöglich ist, alle Aufteilungen in Betracht zu ziehen, und da stattdessen immer nur eine Auswahl überprüft werden kann, sind die Prinzipien der Auswahl, so etwaswie eine implizite Vermutung über die Struktur der Daten. Von der schliesslich gefundenen Aufteilung weiss man nie, ob sie das absolute Optimum ist. Wenn die "implizite Hypothese", die der Suchstrategie zugrunde lag, unangemessen war, stellt sie unter Umständen sogar eine starke Verzerrung der mit den Daten gegebenen Relationen dar.

Wir müssen uns hier wiederum wegen der Zielsetzung der Arbeit in unseren Ausführungen zu Clusterfindungsprozeduren, ihrer Effizienz und ihrer Beziehung zur Datenstruktur auf Andeutungen beschränken, um schneller zum Ergebnis dieser Überlegungen zu gelangen.

Die vorliegenden Clustermethoden unterscheiden sich im Umfang der Vergleichsoperationen, die sie durchführen. Je näher ein Verfahren dem Ideal der vollständigen Durchsicht aller Aufteilungsmöglichkeiten kommt, desto sicherer trifft es auf die absolut beste Lösung. Andererseits ist natürlich der Rechenaufwand entsprechend hoch und schränkt seine Anwendung auf unverhältnismässig kleine Mengen ein.

Ein Extrem in dieser Hinsicht ist das von EDWARDS und CAVALLI-SFORZA (1965) konzipierte Verfahren. Es kommt dem Ideal der Berücksichtigung aller möglichen Aufteilungen fast so nahe wie die unmittelbar nächstliegende Strategie, nur für ausgewählte ks alle Aufteilungen zu inspizieren. Demgegenüber ergibt sich nämlich schon eine erhebliche Ersparnis, wenn man, wie sie vorschlagen, erst einmal die optimale Zweiteilung der Gesamtmenge vornimmt, um dann in rekursiver Fortführung dieses Prinzips die dabei entstehenden Teilmengen weiter zu zweiteilen, bis entweder eine angestrebte k-Teilung oder ein vorgegebener Kriteriumswert erreicht ist. Im Fall, dass die Teilmengen etwa gleich gross sind, vermindert sich der Aufwand beispielsweise für eine so durchgeführte Fünfteilung von 25 Objekten mit $1{,}7 \cdot 10^7$ notwendigen Vergleichen auf ein Millionstel gegenüber der direkten Fünfteilung und bietet den Vorteil, zugleich Aufschluss über gute Zwei-, Drei- und Vierteilungen zu geben. Demgegenüber wiegt der Mangel gering, nun nicht mehr sicher die absolut optimalen Aufteilungen in Händen zu halten, weil eine deutlich in den Daten vorhandene Clusterstruktur sich in jedem Fall in den Aufteilungen widerspiegelt.

Der bei dieser Methode erforderliche Rechenaufwand bewegt sich freilich, selbst für dies bescheiden dimensionierte Klassifikationsproblem, in einer kaum mehr zu bewältigenden Grössenordnung: GOWER (1967) schätzt die Rechenzeit schon für 21 Objekte auf mehr als 100 Stunden bei einer Zugriffszeit von 5 µ-Sekunden.

Jedenfalls haben wir mit der Form des Ergebnisses im Prinzip schon eine Lösung für die eingangs formulierten Probleme vor uns: weder landen wir auf der Suche nach der optimalen Aufteilung bei der trivialen mit k = n, noch sind wir zu einer Vorentscheidung über die gewünschte Anzahl der Klassen oder den Kriteriumswert gezwungen. Dafür existiert gewöhnlich auch gar keine vernünftige Basis, weil erst Einblick in die besondere Struktur der Daten nachträglich darüber Aufschluss geben kann, welche Aufteilung unter den gegebenen Umständen als zweckmässig oder "natürlich" anzusehen ist.

Derartige im voraus getroffene Festlegungen erübrigen sich, weil wir, ausgehend von k = 1 über 2, 3 usf. einen Überblick über beliebig feine Aufteilungen erhalten. Die Klassifikationen unterliegen wegen der Art des Algorithmus lediglich der Restriktion "geschachtelt" zu sein: Die "nächstfeinere" Aufteilung kommt immer dadurch zustande, dass eine Klasse zweigeteilt wird. Zusammen bilden die Aufteilungen, was man ein hierarchisches Klassifikationssystem (Hierarchical Clustering Scheme, HCS) nennt. Glücklicherweise besteht die Möglichkeit, diese Vorzüge auch ohne die mit rechnerischer Aufwendigkeit verbundenen Beschränkungen zu realisieren. Wiederum ist die Hierarchizität der zu bildenden Aufteilungen die Grundlage für die Ersparnis gegenüber dem systematischen Vergleich aller möglichen Aufteilungen. Während jedoch die Methode von EDWARDS und CAVALLI-SFORZA divisiv verfährt, beschreiten die von uns favorisierten und verwendeten Clustertechniken den genau umgekehrten Weg: Die erste Klassifikation ist die triviale Aufteilung in die n Objekte. Durch Vereinigung der zwei ähnlichsten bzw. einander nächsten, wird die zweite Aufteilung erzeugt. Die rekursive Anwendung der Regel, dass die nächst gröbere Aufteilung jeweils durch Zusammenfügen der beiden einander ähnlichsten Cluster/Objekte zu bilden ist, endet nach höchstens n - 1 Schritten mit der Vereinigung aller Objekte.

Die Anzahl der bei diesem Vorgehen benötigen Paarvergleiche ist mit

$$(3) \qquad 2n^2 - 9n/2 \qquad \text{(ANDERBERG , 1973)}$$

ganz erstaunlich gering, bei 100 Objekten sind es zum Beispiel 19550, und die Rechenzeit liegt im Bereich von einer Minute.

## 2.5 Definition des HCS

Wir definieren zunächst, wann eine Aufteilung $P_i$ als "Verfeinerung" einer Aufteilung $P_j$ zu bezeichnen ist.

Definition 2:

Die Ordnungsrelation '$\sqsubset$ ', d.h. ' ist eine Verfeinerung von' ist wie folgt definiert:

$P_i \sqsubset P_j$, wenn es für jede Klasse $c_j^{(i)}$ aus $P_i$ eine Klasse $c_l^{(k)}$ in $P_k$ gibt, in der sie vollständig enthalten ist;
d.h. $c_j^{(i)} \subseteq c_l^{(k)}$. Und für mindestens ein j $c_j^{(k)} \subset c_l^{(k)}$.

Ein Hierarchisches Clustersystem (HCS) läßt sich nun wie folgt definieren:

Definition 3:

H1 $P_0$ ist die triviale Aufteilung auf n Klassen mit jeweils einem Objekt

H2 $P_I$ ist die triviale Aufteilung, die aus einer einzigen Klasse für alle Objekte besteht. Dabei ist $I \leq n-1$.

H3 $P_k \sqsubset P_{k+1}$

Und in einigen Fällen auch:

H4 Es gibt kein $P_x$ mit $P_k \sqsubset P_x \sqsubset P_{k+1}$.

Die zu den einzelnen Aufteilungen eines HCS gehörenden Klassen $c_j^{(i)}$ heissen "Cluster".

Damit der agglomarative Clusterungsprozess in der oben beschriebenen Weise vonstatten gehen kann, muss das zunächst nur für Objektpaare definierte Distanzmass auf die Beziehung zwischen zwei Clustern verallgemeinert werden. Verschiedene Möglichkeiten werden hier praktiziert: Man kann den Abstand der beiden nächstliegenden Objekte aus zwei Clustern zugrundelegen oder auch gerade den der entferntesten. Man kann die durchschnittliche Distanz aller Paare von Objekten berechnen, bei denen jeweils eins dem einen und eins dem anderen Cluster angehört. Man kann zusätzlich eine Gewichtung einführen, um den Einfluss der unterschiedlichen Grössen der zu vereinigenden Cluster auszuschalten usf.. Jede dieser Möglichkeiten ist zugleich eine weitere Spezifikation von 'Cluster' und hat mit Hinblick auf eine bestimmte Datenstruktur bzw. ein Untersuchungsziel einen Sinn. Eine grundsätzliche Bevorzugung e i n e r Methode kann es daher nicht geben. Lediglich die unterschiedlich starken Annahmen über die Skalenqualität oder die Tatsache, dass sie im unterschiedlichen Ausmass von den mit der Matrix vorliegenden Informationen Gebrauch machen, sind unabhängig davon zu berücksichtigen.

Die Bedeutung des hierarchischen Klassifikationssystems liegt aber nicht allein darin, dass es mit dem agglomerativen Algorithmus so ungemein einfach zu erzeugen ist. Ein besonderer Vorteil der Hierarchizität ist, dass sie die Darstellung der Aufteilungen als Baumgraph ermöglicht. Dabei geht keine Information verloren, denn es besteht eine umkehrbar-eindeutige Abbildung (Isomorphismus) zwischen der geordneten Menge der Aufteilungen, die ein HCS bilden und einem Baumgraphen.

Es gibt verschiedene Arten von Baumgraphen. Im Unterschied zu den sonst recht ähnlichen Baumgraphen der Psycholinguistik (CHOMSKY und MILLER ,1963) und der Soziologie (FRIEDELL, 1967), die ebenfalls "rooted trees", Bäume mit einem Hauptstamm sind, werden hier nur die Endknoten benannt und zwar mit den klassifizierten Objekten, während die nicht-terminalen Knoten oder

Gabelungspunkte numerische Werte erhalten. Diese Werte spiegeln den Ablauf des Agglomerationsprozesses wider.

Im einfachsten Fall sind es natürlich Zahlen entsprechend der Reihenfolge, in der die Elemente zusammengefügt wurden. Ein Knoten erhält die Zahl $a_j = i$ wenn sämtliche unter ihm befindlichen Elemente sich bei der Aufteilung $P_i$ in derselben Klasse befinden und es kein $P_l$ gibt mit $P_l \sqsubset P_i$, für das dies ebenfalls gilt. Die zweite Möglichkeit besteht darin, die a als jene reellen Zahlen zu bestimmen, die jeweils durch den Klassifikationsalgorithmus als Distanz* der durch diesen Gabelungspunkt verbundenen Cluster errechnet worden sind.

Für jedes Objektpaar (x,y) gibt es eine Menge von Knoten, unter denen x und y zusammengefasst sind. a(x,y) sei nun der kleinste Wert eines dieser Knoten. a(x,y) ist also der Wert jenes niedrigsten Knoten, unter dem x und y erstmals zusammengefasst sind.

Definition 4: $D_{xy} = df\ a(x,y)$,

Man erhält so Distanzen auf der Objektmenge mit interessanten Eigenschaften. Die $D_{ij}$ bilden nämlich eine sogenannte Ultrametik (JOHNSON, 1967; JARDINE und SIBSON, 1971).

Definition 5: Die Bedingungen U1 - U4 definieren eine Ultrametrik.

Wir können uns ihre Charakteristik leicht vergegenwärtigen. Da zur Definition einer Aufteilung gehört, dass jedes Objekt x nur einmal vorkommt und x von Anfang an, also bereits in der Aufteilung $P_o$ "mit sich" in derselben Klasse ist, ist

U1 $\qquad D_{xx} = a_o = 0.$

* Bei den Ausgangsdaten kann es sich natürlich auch um Ähnlichkeiten handeln. Für das Folgende gilt dann Entsprechendes.

Die strengere, Identität genannte Bedingung, muss jedoch nicht notwendig erfüllt sein:

U2 $\quad D_{xy} = 0 \Leftrightarrow x = y.$

Dagegen ist stets Symmetrie gegeben:

U3 $\quad D_{xy} = D_{yx}$

denn die Definition von $D_{xy}$ als dem Wert (a,y) jenes Knotens, unter dem x und y erstmals zusammengefasst werden, sichert die Symmetrie, da eine Zusammenfassung von x mit y dasselbe ist wie eine Zusammenfassung von y mit x.

Hauptmerkmal ultrametrischer Systeme ist schliesslich die verschärfte Fassung der Dreiecksungleichung. Statt nur

$d_{xy} \leq d_{xz} + d_{yz}$ wie bei Metriken allgemein, gilt zusätzlich

U4 $\quad D_{xy} \leq \text{Max}\{D_{xz}, D_{yz}\}.$

Man kann sich das so klarmachen: x,y und z seien irgend drei Objekte aus einem HCS mit

(4) $\quad D_{xy} = a_j \quad \text{und} \quad D_{yz} = a_k,$

d.h. x und y sind erstmals in derselben Klasse bei der Aufteilung $P_j$, und y und z bei $P_k$. Wegen des beiden Paaren gemeinsamen Objekts y und der Hierarchizität ist eine der beiden Klassen in der anderen enthalten. Man kann nun zwei Fälle unterscheiden:

1. $a_j$ und $a_k$ sind gleich. Das bedeutet aber nichts anderes, als dass die beiden Paare in derselben Aufteilung $P_j = P_k$ erstmals zusammenkommen. Dasselbe gilt dann notwendig auch für x und z.

Damit ist

$$(5) \qquad D_{xy} = D_{yz} = D_{xz}$$

und U4 ist erfüllt.

2. Eine von beiden Distanzen ist grösser. Angenommen dies sei $D_{xy}$ :

$$(6) \quad D_{xy} > D_{yz} \Rightarrow a_j > a_k .$$

Da bei der Aufteilung $P_j$ sowohl x und y wie auch y und z vereinigt sind, sind dort auch x und z in derselben Klasse und ist daher auf jeden Fall

$$(7) \quad D_{xz} \nless a_j \quad \text{d.h.} \quad D_{xz} \leq a_j .$$

$D_{xz} < a_j$ können wir aber leicht ausschliessen, weil das zu Widersprüchen führen würde. Es folgt nämlich aus

$$(8) \quad D_{xz} < a_j \ \vee \ D_{yz} < a_j \Rightarrow D_{xy} < a_j$$

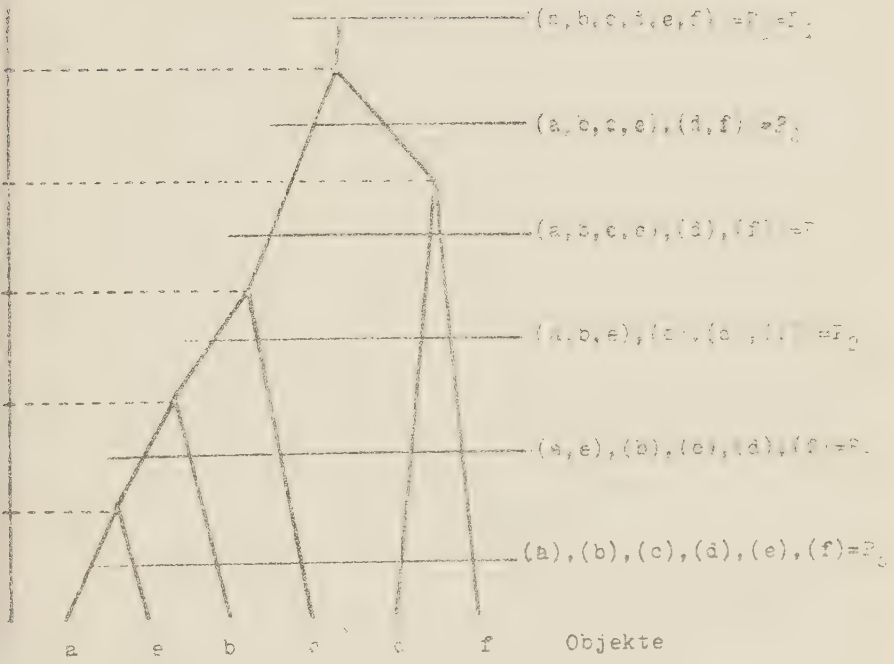
Das aber steht im Gegensatz zu (6). Bleibt also für (7) nur

$$(9) \quad D_{xz} = a_j \quad \text{d.h. aber} \quad D_{xz} = D_{xy} > D_{yz}$$

und damit ist wiederum U4 erfüllt.

Fassen wir zusammen: Die Distanzen jedes Tripels von Objekten aus einem HCS genügen der strengen, ultrametrischen Fassung der Dreiecksungleichung. Das bedeutet, dass entweder alle drei gleich sind - sie geometrisch gesehen ein Gleichseitiges Dreieck bilden - oder dass die kürzere von ihnen die Basis eines Gleichschenkligen Dreiecks ist.

Die mit einem HCS gegebene Datenstruktur lässt sich also auf drei äusserlich sehr verschiedene aber tatsächlich äquivalente Weisen repräsentieren: als geordnete Menge von Aufteilungen, als Baumgraph (ranked tree) und als ultrametrische Datenmatrix. Ihre wechselseitige Austauschbarkeit bildet die Grundlage für die im folgenden Kapitel zu behandelnden Klassifikationsmethoden: Empirisch gegebene, normalerweise nicht-ultrametrische Distanzmatrizen werden - möglichst ohne den Daten Gewalt anzutun - durch ein Hierarchisches Clusterverfahren in ein ultrametrisches System transformiert. Davon ausgehend ist leicht der entsprechende Baumgraph zu konstruieren. "Gute" Aufteilungen der Objektmenge erhält man nun als waagerechte Schnitte durch den Baum. Wünsche über die Anzahl der Klassen oder Schranken hinsichtlich ihrer 'Kohärenz' lassen sich durch die Höhe des Schnittes berücksichtigen. Siehe Tafel 1.




| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | 2 | 3 | 3 | 1 | 3 |
| b | | | 3 | 1 | 2 | 3 |
| c | | | | 3 | 1 | 2 |
| d | | | | | 1 | 4 |
| e | | | | | | 2 |
| f | | | | | | |

## 3. DIE ZUR KLASSIFIKATION VERWENDETEN METHODEN

Wegen der oben dargelegten Vorzüge haben wir ausschliesslich agglomerative, hierarchische Clusteranalysen angewendet. Eine Auswahl von fünf "klassischen" Verfahren erlaubt bei relativ bescheidenem Rechenaufwand den Vergleich von deutlich verschiedenen Clusterkonzeptionen, die aus ihrem Rationale leicht interpretierbar, bereits vielfach analysiert und deshalb in ihrer spezifischen Funktion verstanden waren.

Ihre ökonomische Einsetzbarkeit ist den Bemühungen von LANCE und WILLIAMS (1967) und WISHART (1969) um ihre formale Integration zu verdanken. Ihnen gelang es, einen einheitlichen, agglomarativen Algorithmus für die bis heute wichtigsten Methoden zu entwickeln. Die rechnerische Realisation des "Agglomerierens" besteht, wie auf S. 24 schon angedeutet wurde, in der schrittweisen Reduktion der Distanzmatrix durch Fusion der beiden nächstliegenden Objekte/Cluster. Kernstück des allgemeinen Algorithmus ist dementsprechend die den fünf Methoden gemeinsame Grundgleichung zur Bestimmung der Distanz zwischen einem neugebildeten Cluster (xy) zu jedem der übrigen Objekte/Cluster auf der Basis der Distanzen $d_{xz}$ und $d_{yz}$. Dieser Vorgang ist, wie wir gezeigt haben, gleichbedeutend mit der ultrametrischen Transformation der Distanzmatrix. Durch vier Parameter $a_1$, $a_2$, b und g wird die Grundgleichung

$$(10) \qquad d^2_{(xy)z} = a_1 d^2_{xz} + a_2 d^2_{yz} + b d^2_{xy} + g|d^2_{xz} - d^2_{yz}|$$

zur Bestimmungsgleichung für die fünf bekannten Clustermethoden spezifiziert.

Die vor allem durch JOHNSON (1967) in der Psychologie bekanntgewordenen Methoden 'single-linkage-' und 'complete-linkage-analysis' oder wie sie bei ihm heissen Minimum-und Maximum-Methode zeichnen sich dadurch aus, hinsichtlich der Skalenqualität wenig anspruchvoll zu sein, was der üblichen Messge-

nauigkeit in den Sozialwissenschaften entgegenkommt; sie setzen lediglich ordinalskalierte Daten voraus.

## 3.1 Die Minimum-Methode

Die Namen 'single-linkage' und 'Minimum-Methode' kennzeichnen die Fusionsstrategie dieses Verfahrens. Vereinigt werden nämlich jeweils diejenigen Cluster x und y, die über die kürzeste Distanz zwischen einem Objekt aus x und einem aus y miteinander verbunden sind. (Ausschlaggebend ist also der Wert einer einzigen Distanz). Entsprechend wird nach der Zusammenfassung von x und y die Distanz des auf diese Weise neu gebildeten Clusters (xy) zu jedem anderen z festgesetzt als

Definition 6:

$$d_{(xy)z} = \text{Min}\,(d_{xz}, d_{yz}).$$

oder, was dasselbe ist, die Grundgleichung mit den Parametern

$$a_1 = a_2 = 1/2\ ; \quad b = 0\ ; \quad g = -1/2$$

versehen.

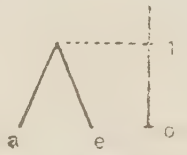
Das Kriterium der Minimum-Methode ist die Minimierung der Maximaldistanz $d_{ij}$ in der in dieser Hinsicht günstigsten (Reihen-)Folge der zusammengefassten Objekte. Cluster, die nach dieser Methode gebildet werden, können also sehr weit entfernte Objekte einschliessen. Bestimmend für ihr Zustandekommen ist lediglich ihre dichte Verkettung oder enge Verbundenheit (engl. 'connectedness') über nicht weiter als $d_{ij}$ voneinander entfernte Glieder. $d_{ij}$, es erscheint im Baumgraph als Höhe des Knotens, unter dem sich die zum Cluster vereinigten Objekte befinden, gilt als "Wert dieser Cluster" und ist im eben dargelegten Sinn zu interpretieren, wenn die Klassifikation durch die Minimum-Metho-

de erfolgte. Zum andern besagt es, dass die Cluster mindestens durch den Abstand $d_{ij}$ von allen übrigen Clustern dieser Stufe getrennt sind, sich aber sonst jeder räumlichen Verteilung der Objekte anpassen können. Das ist ein wichtiger Unterschied zu allen anderen Methoden, die zu maximal kompakten und konvexen Clustern tendieren, ohne jedoch deren räumliche Separation zu berücksichtigen. Interessante Varianten dieses Ansatzes von JARDINE und SIBSON (1968), COLE und WISHART (1970) und vor allem LING (1973) beruhen auf der Verallgemeinerung die Anzahl der notwendigen Verbindungen betreffend. Statt die Vereinigung von einem 'single-link' abhängig zu machen, wird eine geeignete Anzahl von 'k-linkages' zum Kriterium der Fusion.

Numerisches Beispiel für die Durchführung des Hierarchischen Clustering nach der Minimummethode.

Die Ausgangsmatrix enthält, wie das bei empirischen Daten der Fall ist, nicht-ultrametrische Distanzen. Der Übersichtlichkeit halber wurden sie so gewählt, dass sie gleichzeitig als Ränge verstanden werden können.

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | 2 | 3 | 5 | 1 | 7 |
| b | | | 3 | 5 | 4 | 6 |
| c | | | | 7 | 3 | 6 |
| d | | | | | 6 | 4 |
| e | | | | | | 5 |
| f | | | | | | |



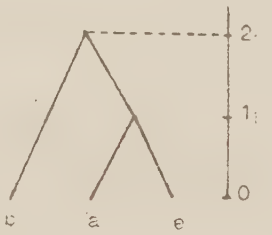

Erläuterung siehe nächste Seite.

Tafel 2: Ausgangsmatrix und Teil des Baumes aus erster Fusion

Die kleinste Distanz in der Matrix ist $d_{ae} = 1$. Daher werden auf der ersten Stufe a und e zu einem Cluster zusammengefasst. Dem entspricht als erstes Stück des Baumgraphen eine "Astgabel" mit a und e als Enden und der Knotenhöhe 1.

Nun gilt es nach Definition 6 die erste reduzierte Matrix zu berechnen. Es ergibt sich für $d_{(ae)b} = \text{Min}(d_{ab}, d_{eb}) = \text{Min}(2,4) = 2$ und entsprechend $d_{(ae)c} = 3, d_{(ae)d} = 5$ und $d_{(ae)f} = 5$ (In den Tafeln mit '* ')
Die übrigen Distanzen bleiben unverändert.

| | (ae) | b | c | d | f |
|---|---|---|---|---|---|
| (ae) | | 2 | 3* | 5* | 5* |
| b | | | 3 | 5 | 6 |
| c | | | | 7 | 6 |
| d | | | | | 4 |
| f | | | | | |

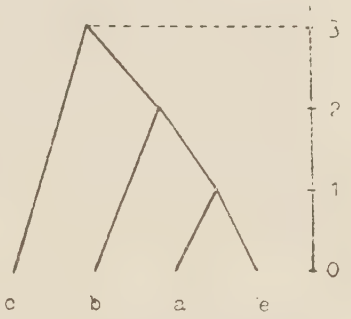



Tafel 3: 1. reduzierte Matrix und dazugehöriger Baumgraph

Die kleinste Distanz ist $d_{(ae)} = 2$. Also wird b dem schon bestehenden Cluster (ae) zugeschlagen. Der Baumgraph bildet die Fusion von (ae) mit b auf Stufe 2 durch eine Gabelung dieser Höhe ab.

Für die 2. reduzierte Matrix werden auf der Basis der 1. reduzierten die Distanzen von (abe) nach c, d und f ermittelt:

| | (abe) | c | d | f |
|---|---|---|---|---|
| (abe) | | 3° | 5° | 5° |
| c | | | 7 | 6 |
| d | | | | 4 |
| f | | | | |

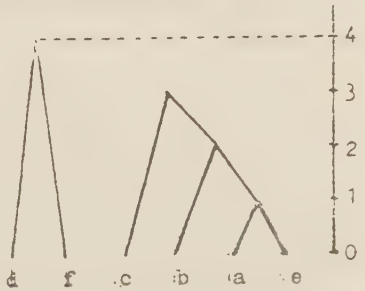



Tafel 4: 2. reduzierte Matrix und dazugehöriger Baumgraph

Das bestehende Cluster wird um c erweitert und der Baum entsprechend ergänzt.

In der 2. reduzierten Matrix ist $d_{df} = 4$ die kleinste Distanz. Es sind für die folgende Matrix also $d_{(abce)d}$ und $d_{(abce)f}$ zu ermitteln.

| | (abce) | d | f |
|---|---|---|---|
| (abce) | | 5° | 5° |
| d | | | 4 |
| f | | | |




Tafel 5: 3. reduzierte Matrix und dazugehöriger Baumgraph

In dieser Matrix ist $d_{df} = 4$ die kleinste Distanz.Dadurch entsteht ein neues Cluster mit d und f als einzigen Elementen, dem ein neuer Ast entspricht mit dem Gabelungspunkt auf der Höhe 4.

Aus der 3. reduzierten Matrix ist nun noch die Distanz zwischen den beiden alle Objekte einschliessenden Clustern zu bestimmen, dann ist die Hierarchische Klassifikation nach der Minimum-Methode vollzogen. Wie man unschwer erkennt, beträgt der Wert $d_{(abce)(df)} = 5$ und damit vereinigen sich auch die beiden Hauptäste auf dieser Höhe zum "Stamm" des Baumes. Seine graphische Darstellung an dieser Stelle erübrigt sich, weil er identisch mit dem Baum der Tafel 1 ist. Dort findet sich auch die Folge der Aufteilungen und die ultrametrische Matrix, die ebenso wie der Graph das Ergebnis dieser Klassifikation ausmachen.

## 3.2 Die Maximum-Methode

Die Maximum-Methode zielt, bildlich gesprochen - darauf ab, möglichst "kompakte" Cluster zu erzeugen. Für die Entscheidung darüber, welche Cluster jeweils vereinigt werden sollen, vergleicht der Algorithmus alle Paare von Clustern der betreffenden Stufe hinsichtlich der jeweils grössten Distanz zwischen den Objekten des einen und des andern Clusters. Zusammengefasst werden diejenigen Cluster, bei denen diese Distanz am kleinsten ist. Auf diese Weise wird erreicht, dass die maximale Distanz innerhalb - der "Wert des Clusters" bei dieser Methode - stets minimal ist.

Die Funktion der Grundgleichung ist es, nach Fusion der im obigen Sinne "nächsten" Cluster x und y die Distanz von (xy) zu jedem anderen Cluster z wiederum als Maximum der Distanzen von x und y zu z , $\mathrm{Max}(d_{xz}, d_{yz})$, festzusetzen. Das geschieht durch die Grundgleichung mit den Parametern

$$a_1 = a_2 = 1/2 \; ; \; b = 0 \; ; \; g = 1/2.$$

Positiv hervorgehoben zu werden verdient an Maximum- und Minimum-Methode, dass sie eine einfache und anschauliche Interpretation für den "Wert" und damit auch die Bedeutung

des Clusters haben. Ihre rechnerische Durchführung ist vergleichsweise sehr einfach,und wegen der Invarianz der verwendeten Operationen gegenüber monotonen Transformationen der Distanzen brauchen die Daten nur Ordinalskalenniveau aufzuweisen. Diese Vorzüge sind jedoch begleitet von einem unter Umständen nicht unerheblichen Nachteil: Kriterium für die Fusion von zwei Clustern ist ein einziger Distanzwert. Im Interesse einer grösseren "Robustheit" des Ergebnisses gegenüber Fehlervarianz (s. JOHNSON, 1967, S. 52) sollte sich diese Entscheidung aber auf möglichst viel Information stützen.

## 3.3 Die Methode "Gruppendurchschnitt"

(n. 'group average' von LANCE und WILLIAMS, 1967)

Es liegt nahe, statt sich bei der Bestimmung der Distanz von Clustern nur auf einen der Extremwerte zu stützen, z. B. irgendeine Art von Mittelwert zu verwenden. Der Median der Zwischen-Cluster-Distanzen teilt zwar mit Minimum und Maximum die Eigenschaft der monotonen Invarianz, der Agglomerationsprozess hat sich aber, weil der Median nicht notwendig monoton mit der hierarchischen Stufe wächst, häufig als undurchführbar erwiesen. Die folgenden drei Methoden machen sämtlich von anderen Mittelungsoperationen Gebrauch und setzen deshalb mindestens intervallskalierte Daten voraus.

Prototyp einer Reihe ähnlicher Verfahren ist die Methode "Gruppendurchschnitt", bei der die Distanz zwischen zwei Clustern x und y einfach als das arithmetische Mittel aller Zwischen-Cluster-Distanzen ihrer Objekte definiert ist:

$$\frac{\sum_{i \in x,\ j \in y} d_{ij}}{n_x \cdot n_y}$$

Die Grundgleichung mit den Parametern

$$a_1 = n_x/n_{(xy)} \; ; \; a_2 = n_y/n_{(xy)} \; ; \; b = g = 0$$

ergibt die Distanz des aus der Vereinigung von x und y hervorgegangenen Clusters (xy) zu einem dritten z.

Weil der Durchschnitt grundsätzlich ohne Berücksichtigung der Variabilität eine unvollständige Beschreibung darstellt, ist für die Cluster dieser Methode 'Separation' in irgendeinem Sinne nicht definiert und dementsprechend auch nicht berücksichtigt.

## 3.4 Die Schwerpunktmethode

Die Schwerpunktmethode kann man sich so veranschaulichen, dass jedes Cluster durch seinen Schwerpunkt ('centroid') repräsentiert wird und diejenigen CIuster vereinigt werden, deren Schwerpunkte einander am nächsten sind. Die zur Reduzierten Matrix führende Neubestimmung der Distanzen des gerade gebildeten zu den übrigen Clustern über die Grundgleichung mit den Parametern

$$a_1 = n_x/n_{(xy)}; \; a_2 = n_x/n_{(xy)} \; ; \; b = a_1 a_2 \; ; \; g = 0$$

stellt eine Art Gewichtung durch die jeweilige Anzahl der Objekte in den fusionierten Clustern bzw. durch die Masse ihrer Schwerpunkte dar. Dies hat die - gerade wenn sich eine grössere Zahl identischer Objekte darunter befindet - unerwünschte Wirkung, dass die Distanz des neuen (xy) zu jedem anderen Cluster z stärker an seiner Entfernung vom Schwerpunkt des grösseren der beiden vereinigten Cluster orientiert ist.

Siehe Abb. 2 unten.

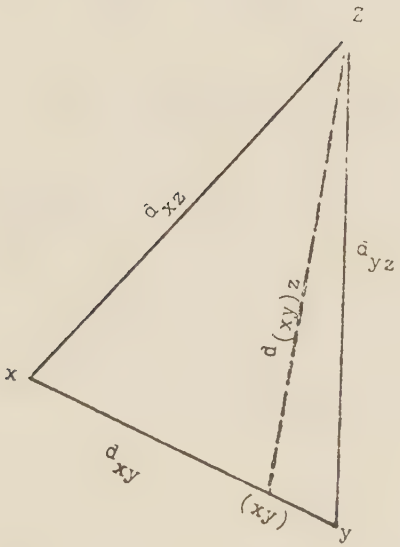



Abb. 2: Geometrische Darstellung der Schwerpunktmethode

x und y werden nach ihrer Vereinigung durch ihren gemeinsamen Schwerpunkt (xy) repräsentiert. Der liegt daher auf der Strecke zwischen beiden und zwar so, dass er sie im Verhältnis ihrer jeweiligen Grösse teilt. Die Distanz zu einem dritten Cluster z ist dann einfach der Abstand seines Schwerpunktes von diesem Punkt (xy).

## 3.5 WARDs Methode

WARD stellte 1963 einen allgemeinen Algorithmus zur Hierarchischen Klassifikation gemäss einem beliebigen Kriterium W vor. Bekannt geworden ist sein Verfahren mit dem von ihm selbst verwendeten Beispiel für W als der Gesamtbinnenvarianz der Cluster einer Stufe.
Im Unterschied zu den beiden vorigen hat die Methode von WARD auch wieder eine einfache und sinnvolle Interpretation eines Clusters - dies allerdings strenggenommen nur, wenn es sich bei den $d_{ij}$ um Euklidische Distanzen handelt. $d_{xy}$ ist dann genau die Varianz des durch die Vereinigung von x und y entstandenen Clusters. Die nächstliegenden und deshalb auf der jeweiligen Stufe zu vereinigenden Cluster sind diejenigen, bei denen die Summe der quadrierten Abweichungen (innerhalb)

Definition 7:

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

g = Anzahl der Klassen
$n_i$ = Anzahl Objekte/ Klasse i,

$$n = \sum_{i=1}^{g} n_i$$

gegenüber dem Zustand vor der Fusion den kleinsten Zuwachs erfährt. Für die Parameter der Grundgleichung ergibt sich dementsprechend:

$$a_1 = (n_z+n_x)/(n_z+n_{(xy)}) \quad ; \quad a_2 = (n_z+n_y)/(n_z+n_{(xy)}) \; ;$$

$$b = - n_z/n_z+n_{(xy)} \; ;$$

Wegen der zwischen ihnen bestehenden Additivitätsbeziehung ist die Minimierung von SSinnerhalb gleichbedeutend mit der Maximierung von SSzwischen. Die Methode von WARD kann als die elaborierteste Technik, die sich in die Grundgleichung fassen lässt, auf diese Weise als Einzige die intuitiven Kriterien Homogenität und Separation gleichzeitig optimieren.

Sie hat den Nachteil, mit den Ausgangsdaten gegebene ultrametrische Tripel nicht notwendig zu erhalten.

Für sie wie für alle anderen agglomerativen, hierarchischen Verfahren gilt, dass wegen der Nicht-Revidierbarkeit von auf früherer Stufe getroffenen Entscheidungen über das Vereinigen von Clustern, die für eine bestimmte Anzahl von Klassen optimale Aufteilung nicht unbedingt gefunden werden muss.

## 4. DIE LOGIK DER KOMBINATION VON SORTIERMETHODE UND H.C.S.

### 4.1 Die Sortiermethode

Eine wegen ihrer methodischen Reflektiertheit und der daraus resultierenden Gegenstandsangemessenheit herausragende Anwendung der Hierarchischen Clusteranalyse veröffentlichte 1969 G. A. MILLER. Die als Eingabeinformation für die Clusteranalyse erforderliche symmetrische Distanzmatrix über die betreffende Objektmenge lieferte ihm eine bis dahin nicht sehr bekannte Skalierungstechnik: die Sortiermethode.

Dies in seiner Durchführung so unkomplizierte Verfahren, verlangt von den Vpn lediglich kategoriale Entscheidungen. Die Objektmenge soll von der Vp so aufgeteilt werden, dass einander ähnliche Objekte sich in derselben Teilmenge zusammenfinden. Der Vp ist freigestellt, wieviele und wie grosse Untergruppen sie bilden will. Dadurch dass jedes Objekt nur einmal vorhanden ist, ergibt sich in jedem Fall eine Aufteilung im Sinne von A1 - A3.

Betrachtet man nun alle Paare von Objekten daraufhin, ob sie sich in einer Aufteilung in derselben Klasse befinden und ordnet ihnen dann die Distanz '0' zu (denn das bedeutet ja, dass die Vp die bestehenden Unterschiede zu vernachlässigen bereit war) und die Distanz '1', wenn sie getrennt sind, und summiert über alle Vpn beziehungsweise deren Aufteilungen, so erhält man eine symmetrische Distanzmatrix. Von ihr nimmt MILLER an, dass sie das implizite Urteil aller Vpn über die quantitativen Relationen zwischen den Objekten widerspiegelt. Zwar waren die Vpn durch die Art der Erhebungsprozedur zur Abgabe von Alles-oder-Nichts-Urteilen, "in einer Klasse vereinigt = ähnlich" und "getrennt = unähnlich", gezwungen, aber dadurch,dass sie in der Wahl von Anzahl und Grösse der Klassen frei waren, konnten **verschiedene** Vpn ihrem Urteil **verschiedene Ähnlichkeits grade** zugrunde legen.

Wurde dasselbe Paar von Objekten von vielen Vpn in einer Klasse vereinigt - so war seine Schlussfolgerung - musste es sich um sehr ähnliche Objekte handeln, geschah das dagegen nur selten, so nahm er an, dass sie nur bei einem sehr weitgefassten Ähnlichkeitsbegriff als zusammengehörig empfunden wurden.

Wie diese Skizze seiner Vorgehensweise zeigt, setzt MILLERs Argumentation voraus, dass es sich bei den Vpn um 'identical subjects' handelt, die, ausgehend von derselben hierarchischen Struktur der Objekte sich nur bezüglich der Ähnlichkeitsanforderung, die sie zum Kriterium des Zusammenfügens machen, graduell unterschieden. Ihre Aufteilungen sollten demnach also nur verschiedenen Schnitten durch denselben Baum entsprechen.

Im Fall von MILLERs psycholinguistischer Fragestellung - ihm ging es um die kognitive Organisation unseres internalisierten Lexikons - war der allgemeinpsychologische Ansatz wohl auch vertretbar, und so konzentrierte er sich auf die folgenden Probleme:

a) Welche lexikalischen Strukturen sind zu erwarten, wenn man sich die Bedeutung von Begriffen aus dem Vorhandensein oder Fehlen elementarer Bedeutungsträger (im folgenden 'Merkmale' genannt) erklärt ? Er stützt sich dabei auf die sogenannte "Feature-Theorie der Sematik" von KATZ und FODOR (1964).

b) Wie würden verschiedene solcher Strukturen durch die Kombination von Sortiermethode und Hierarchischer Klassifikation abgebildet ?

Die Ergebnisse, zu denen MILLER dabei kommt, sind wegen ihres formalen Charakters geeignet, ganz allgemein die Bedingungen und Möglichkeiten dieser Methodenkombination zu erhellen.

Im Interpretationsrahmen dieser linguistischen Theorie er-

scheint das Sortierverhalten der Vpn als Ergebnis der Beachtung bzw. Vernachlässigung von Merkmalen. Begriffe oder allgemein Objekte werden in einer Klasse vereinigt, wenn den Merkmalen, in denen sie sich unterscheiden, keine Beachtung geschenkt wird. Objekte einer Klasse hätten danach dieselben Merkmale gemeinsam. Umgekehrt wird die Trennung von Objekten durch ihre Verteilung auf verschiedene Klassen aus der Berücksichtigung wenigstens eines Merkmals erklärt, bezüglich dessen sie sich unterscheiden. Diese Annahme hat charakteristische Konsequenzen für die Interpretierbarkeit der Aufteilungen je nachdem ob die [illegible] paradigmatische, lineare oder hierarchische Organisationsstruktur aufweisen.

Im Falle einer paradigmatischen Organisation sind alle Objekte durch dieselben Merkmale gekennzeichnet, d.h. Vorhandensein oder Fehlen jedes dieser Merkmale hat für jedes Objekt einen unterscheidenden Sinn. Ein Beispiel mit Verwandschaftsbezeichnungen mag das erläutern.

Die Begriffe 'Vater', 'Mutter', 'Sohn', 'Tochter', werden durch Generation ('+' = Eltern, '-' = Kind) und Geschlecht ('+' = männlich, '-' = weiblich) charakterisiert.

| | $W_h$ Vater | $W_i$ Mutter | $W_j$ Sohn | $W_k$ Tochter |
|---|---|---|---|---|
| $F_1$ Generation | + | + | - | - |
| $F_2$ Geschlecht | + | - | + | - |

Tafel 6: Objekt-Merkmals-Matrix für paradigmatische Organisation

Neben allem, was die vier Begriffe gemeinsam haben mögen, (z.B. dass es sich jeweils um Menschen handelt) ist die Kenntnis von Generation und Geschlecht notwendig und hinreichend für ihre Unterscheidung. Auf der Basis dieses Merkmalsmodells sind nur drei verschiedene Aufteilungen denkbar: Eine Vp, die $F_1$ vernachlässigt, bildet zwei Klassen, eine mit Vater und Sohn und eine mit Mutter und Tochter, dargestellt durch hj/ik. Bei Nichtberücksichtigung von $F_2$ entsteht hi/jk. Bezeichnen wir nun mit n(hi/jk) die Anzahl der Vpn, die die entsprechende Aufteilung vorgenommen haben, und bedeutet $N_{ij}$ die Häufigkeit mit der die Begriffe/Objekte $W_i$ und $W_j$ über alle N Vpn in einer Klasse vereinigt waren, dann ist

Definition 8: $D_{ij} =_{df} N - N_{ij}$

als Distanz von $W_i$ und $W_j$ definiert. Die $D_{ij}$ bilden wie MILLER nachweist, eine Pseudometrik, die sich von einer echten Metrik nur durch die Abschwächung der Identitätsbedingung unterscheidet:

Relation 9: $M_1$ - $M_3$ definieren eine Pseudometrik.

$M_1$ $\quad D_{ij} = 0$ für $i = j$ und $N_{ij} = N$;

$M_2$ $\quad D_{ij} = D_{ji}$ für alle i,j;

$M_3$ $\quad D_{ij} = D_{ik} + D_{jk}$ für alle i,j,k

Vpn, die ein Quadrupel von der Art der vier Verwandtschaftsbezeichnungen zu sortieren hätten, liessen folgendes Ergebnis erwarten:

(11) $N_{hi} = N_{jk} = n(hijk) + n(hi/jk)$

(12) $N_{hj} = N_{ik} = n(hijk) + n(hj/ik)$

(13) $N_{hk} = N_{ij} = n(hijk).$

Da n(hijk) in allen Gleichungen als Summand auftritt, gilt für die Distanzen:

(14) $D_{hi} = D_{jk} < D_{hk} = D_{ij}$ und

(15) $D_{hj} = D_{ik} < D_{hk} = D_{ij}.$

Über das Verhältnis von $D_{hi} = D_{jk}$ und $D_{hj} = D_{ik}$ macht das Modell keine Aussagen. Ein empirischer Unterschied zwischen beiden wäre dem Umfang und der Richtung nach unvorhersagbare Fehlervarianz und theoretisch bedeutungslos. Es ist nun aber leicht zu zeigen, dass, wie auch immer dies Verhältnis ist, die Distanzen nicht die Ultrametriebedingung erfüllen können.

In den beiden folgenden Beweisen setzen wir die Geltung der ultrametrischen Dreiecksungleichung U4 voraus und zeigen, dass daraus Widersprüche mit (14) und (15) resultieren. Angenommen es sei $D_{hi} = D_{jk} = D_{hj} = D_{ik}$. Wegen U4 folgt daraus $D_{ij} \leq D_{hi} = D_{hj}$. Das steht aber im Widerspruch zu (15) mit $D_{hi} < D_{ij}$. Es bleibt noch die Möglichkeit, dass eines der beiden Paare gleicher Distanzen grösser als das andere ist. Nehmen wir an es sei $D_{hi} = D_{jk} < D_{hj} = D_{ik}$. Bei $D_{jk} < D_{ik}$ erwarten wir wegen U4 für $D_{ij} \leq \text{Max}(D_{jk}, D_{ik})$ also $D_{ij} \leq D_{jk}$. Und das steht wiederum im Gegensatz zu (15), wonach $D_{ij} > D_{ik}$ ist.

Eine paradigmatische Beziehung – oft auch 'cross-classification' genannt – wie sie zwischen den vier Objekten besteht, kann also durch ein HCS nicht angemessen repräsentiert werden. Je nach Art des Klassifikationsverfahrens wird die Existenz wenigstens eines der beiden diskriminierenden Merkmale völlig unterschlagen.

Von linearer Organisation spricht MILLER, wenn wie z.B. bei 'Baby', 'Kind', 'Jugendlicher' und 'Erwachsener' die Objekte eine Sequenz bilden. Für jedes Tripel solcher Objekte erwarten wir Distanzen, die der Gleichung

(16) $D_{ij} + D_{jk} = D_{ik}$

genügen.

Als Ergebnis des Sortierverfahrens ist eine solche Relation der Distanzen aber schon durch die Art der Skalierung prinzipiell ausgeschlossen. Was stattdessen in diesem Fall geschieht, soll kurz erläutert werden.

Für die linear geordneten Objekte $W_i$, $W_j$ und $W_k$ wird die Aufteilung ik/j nicht erwartet. Durch Umformung nach Definition 8 erhalten wir daher für

(17) $D_{ij} = n(i/j/k) + n(jk/i)$,

(18) $D_{jk} = n(i/j/k) + n(ij/k)$ und

(19) $D_{ik} = n(i/j/k) + n(ij/k) + n(jk/i)$.

Daraus folgt

$$D_{ij} + D_{jk} = D_{ik} + n(i/j/k).$$

Weil die additive Konstante n(i/j/k) für jedes Tripel aus einer linearen Sequenz einen anderen Wert haben kann, ist den Daten nicht anzusehen, dass eine solche Organisation gegeben ist. Hier versagt also schon das Sortierverfahren bei der Abbildung der zugrundeliegenden Struktur. Erst recht sind derartige Distanzen mit der Ultrametriebedingung unvereinbar. So ist nicht, wie es erforderlich wäre,

$$D_{ik} \leq \text{Max}(D_{ij}, D_{jk}) \tag{20}$$

sondern grösser als jede von ihnen, denn nach (17) und (19) bzw. (20) und (19) ist

$$D_{ik} = D_{ij} + n(ij/k) \quad \text{bzw.} \tag{21}$$

$$D_{ik} = D_{jk} + n(jk/i). \tag{22}$$

Eine hierarchische Organisation unterscheidet sich von der paradigmatischen dadurch, dass nicht alle Objekte durch Werte auf sämtlichen Merkmalen gekennzeichnet sind. Zwischen den Merkmalen besteht eine bestimmte Form der Abhängigkeit: Nur wenn ein bestimmtes Merkmal $F_1$ in einer bestimmten Ausprägung vorliegt, stellt sich überhaupt die Frage nach dem Vorhandensein eines bestimmten anderen Merkmals $F_2$. Man kann auch sagen $F_1$ dominiert $F_2$. MILLER gibt das Beispiel von Lebewesen ($F_1$), die sich in Pflanzen (+) und Tiere (-) gliedern. Nur auf letztere hat es Sinn, das Merkmal $F_2$ mit den Ausprägungen Wirbeltier (+) und Wirbelloses (-) anzuwenden. Siehe Abb.3.

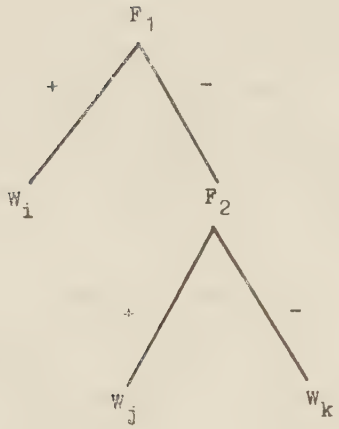



Abb. 3 Hierarchisches System, bei dem das Merkmal $F_2$ von $F_1$ dominiert wird, d.h. $F_2$ nur definiert ist für Objekte, die bezüglich $F_1$ den Wert '-' haben.
(nach MILLER, 1969)

Wenn die Dominanzrelationen zwischen den Merkmalen einen gerichteten Graphen von der Art wie er in Abb. 3 dargestellt ist, bilden, hat das bemerkenswerte Konsequenzen für das Sortieren der Objekte. Die Vernachlässigung eines bestimmten Merkmals hat nämlich zwangsläufig zur Folge, dass auch die von diesem dominierten mit vernachlässigt werden müssen. Wer die Unterscheidungsdimension 'belebt-unbelebt' nicht berücksichtigt, d.h.

Belebtes und Unbelebtes in einer Klasse vereinigt, kann unmöglich hinsichtlich des von ihr dominierten Merkmals 'tierisch-pflanzlich' diskriminieren.
Wann immer i und j in derselben Klasse sind, befindet sich auch k dabei. Die danach zu erwartenden Aufteilungen i/j/k, i/jk und ijk bilden zusammen ein HCS, und andere können bei Zugrundelegung dieser Merkmalsstruktur nicht auftreten.
Das heisst:

(23) $N = n(ijk) + n(i/jk) + n(i/j/k).$

Und es ist

(24) $N_{ij} = n(ijk),$

(25) $N_{jk} = n(ijk) + n(i/jk)$ und

(26) $N_{ik} = n(ijk).$

Diese Grössen stehen in folgender Relation zueinander:

(27) $N \geq N_{jk} \geq N_{ij} = N_{ik} \geq 0.$

Für die entsprechenden Distanzen nach Definition 8 bedeutet das:

(28) $0 \leq D_{jk} \leq D_{ij} = D_{ik} = N.$

Distanzen, die in dieser Beziehung zueinanderstehen, erfüllen aber, wie sich leicht zeigen lässt (JOHNSON, 1967), die ultrametrische Dreiecksungleichung, d.h. bei beliebiger Indizierung gilt

U4 $D_{jk} \leq \mathrm{Max}(D_{ij}, D_{ik}).$

Anders ausgedrückt, wenn die Vpn beim Sortieren der Objekte von derselben hierarchischen Merkmalsstruktur ausgehen, bilden ihre Aufteilungen ein HCS. Dieses HCS lässt sich durch eine Hierarchische Clusteranalyse von der im vorigen Kapitel beschriebenen Art leicht anhand der Distanzmatrix, die sich nach A'1 und A'2 aus den Aufteilungen errechnen lässt, rekonstruieren. Im Prinzip ist auch ein gewisses Ausmass an 'noise' also kleinen, unsystematischen Abweichungen von dem 'allgemeinen' HCS kein Problem, weder bei der rechnerischen Durchführung, noch für die Interpretation.

## 4.2 Das Problem der inhomogenen Versuchspersonengruppe

Die von MILLER selbst nicht problematisierte Voraussetzung seiner Arbeit, die Annahme, dass seine Vpn von derselben lexikalischen Struktur ausgehend, sich nur in Bezug auf ein persönliches Homogenitätskriterium bei der Klassenbildung unterschieden, erwies sich bei genauerem Hinsehen als fragwürdig. In einer Nachfolgeuntersuchung im deutschen Sprachraum (ROSSBACH, 1969) stellte sich heraus, dass die Instruktion, Klassen ähnlicher Begriffe zu bilden, von den Vpn teils denotativ, teils konnotativ verstanden wurde. Wenn daraus für die anschliessende Analyse der Daten trotzdem keine Schwierigkeiten resultierten, so deshalb, weil nur die ins Auge gefasste denotative Auffassung Allgemeines aufdeckte, während die konnotative Auffassung als zu personspezifische Reaktion keinen systematischen Effekt hervorrief.
Aber diese Beobachtung gab den Anstoss zu Zweifeln an der Zweckmässigkeit der selbstverständlichen Zusammenfassung sämtlicher Aufteilungen in e i n e r Distanzmatrix zur anschliessenden Clusteranalyse. Der mit der Reduktion der Sortingdaten auf die Form der gemeinsamen Matrix verbundene Informationsverlust ist umso bedenklicher, als

danach über die Angemessenheit der Zusammenfassung gar nicht mehr geurteilt werden kann. Es ist, als würde man den Mittelwert zur Charakterisierung einer Population verwenden, ohne sich um die dazugehörige Verteilung zu kümmern.

Es besteht also folgendes Dilemma: Einerseits ist die Kombination von Sortiermethode und Hierarchischer Clusteranalyse ein zur Erfassung und Darstellung kognitiver Gehalte vermutlich besonders brauchbares Instrument. Dies ist umso höher zu gewichten, als für die Art der abzubildenden Strukturen geeignete vergleichbare Methoden nicht existieren. Andererseits setzt aber nun diese Verfahrenskombination unabdingbar Aufteilungen voraus, die von derselben Grundstruktur ausgehen, weil sonst ihr Ergebnis,das HCS,keine sinnvolle Interpretation der Ausgangsinformation darstellt.

Der Weg zu einer Lösung dieses Dilemmas, wie wir ihn mit dieser Arbeit vorschlagen, verläuft folgendermassen: Die stillschweigende Prämisse, es mit 'identical subjects' zu tun zu haben, wird als in vielen Fällen unrealistisch fallengelassen und demgegenüber ein differentieller Ansatz fruchtbar zu machen versucht. Statt davon auszugehen, dass die Gruppe der Vpn bzw. deren Aufteilungen homogen hinsichtlich eines HCS sind, wird das zu einer empirischen Frage erklärt. Die Aufteilungen sind demnach daraufhin zu untersuchen, ob

a) sie für eine Zusammenfassung hinreichend homogen sind oder
b) sich unter ihnen solche homogenen Teilmengen befinden oder
c) sie sich vielleicht aus homogenen Untergruppen zusammensetzen.

Die Konkretisierung unseres Lösungsansatzes nimmt ihren Ausgang beim Konzept der 'strukturellen Homogenität' von Aufteilungen. Da zwischen einem HCS und einer bestimmten Menge von Aufteilungen eine ein-eindeutige Abbildungsrelation besteht, liegt es nahe, Aufteilungen aus derselben derartigen Menge als 'strukturell homogen' zu definieren, bzw. im reduzierten Fall zweier Aufteilungen vor ihrer 'Kompatibilität' zu sprechen, wenn für sie '$\subset$' (d.h. 'ist eine Verfeinerung von', s. Definition 2 S. 25) definiert ist. Damit gilt für jedes Paar aus einer Menge strukturell homogener Aufteilungen, dass es kompatibel ist und umgekehrt.

Diese Beziehung würde schon die Möglichkeit zur Identifizierung der zur Zusammenfassung geeigneten Aufteilungen bieten. Es käme nur noch darauf an, diejenigen Aufteilungen zu isolieren, die jeweils untereinander kompatibel sind. Die Funktion, entsprechende Gruppen ausfindig zu machen, erfüllt ideal das Verfahren zur Ermittlung von 'Cliquen' (RATTINGER, 1973).
Aus praktischen Gründen ist diese theoretisch faszinierend klare Lösung in dieser Form jedoch nicht zu verwirklichen. Denn selbst dort, wo die Annahme von 'identical subjects' so wenig umstritten ist, wie bei der psycholinguistischen Untersuchung von MILLER, sind nicht sämtliche Aufteilungen untereinander kompatibel. Kompatibilität als kategoriale Eigenschaft, die entweder zwischen Aufteilungen besteht oder nicht besteht wird dem empirischen Phänomen der - letztlich zu vernachlässigenden - noise-Komponente bei jeder Menge von Aufteilungen nicht gerecht. Die experimentelle Realität legt es also nahe, das Konzept flexibler auszudeuten und von der Alles-oder-Nichts-Entscheidung über Vorliegen oder Fehlen der Kompatibilität zu einer quantitativen Beschreibung der 'strukturellen Ähnlichkeit' zweier Aufteilungen überzugehen.

Diese "Aufweichung" des Kompatibilitätskriteriums ist umso weniger problematisch, als ja auch die Verfahren zum Hierarchischen Clustering, mit denen die Distanzmatrizen der Gruppe weiterverarbeitet werden, vollständige Kompatibilität der Aufteilung (oder, was damit gleichbedeutend ist, Ultrametrie der Distanzen) gar nicht streng voraussetzen. Es muss lediglich unterstellt werden können, dass Abweichungen davon nur zufälliger Natur sind. Bei dieser Annahme kann die Herstellung der Ultrametrie durch das Hierarchische Klassifikationsverfahren als Reinigung der Daten von solchen unsystematischen Einflüssen aufgefasst werden.

Es liegt nahe, strukturelle Ähnlichkeit durch ein geeignetes Distanzmass für Aufteilungen zu explizieren. Was für diesen Zweck "geeignet" heisst, bedarf weiterer Ausführungen. Soviel aber ist klar: Die Identifikation homogener Teilmengen, verstanden als Suche nach Klassen mit hoher innerer Kohäsion (geringer Distanz der zusammengefassten Objekte untereinander) und deutlicher Separation von anderen Klassen (grössere Distanzen zu Objekten aus anderen Klassen) führt wiederum auf die typisch taxonomische Problemstellung .

Mit einigem Recht könnte man eine derartige Umkehrung von Klassifizierungsgegenstand und Informationsbasis als "inverse" Clusteranalyse bezeichnen. Denn bei der hier ins Auge gefassten Klassifikation von Aufteilungen werden ja nicht Distanzen von Objekten, wie sie sich nach Definition 3 aus der Zusammenfassung der Aufteilungen über alle Vpn ergeben, zur Konstruktion eines HCS der Objekte benutzt, sondern zunächst werden Aufteilungen bzw. Vpn anhand ihrer Distanzen klassifiziert. Dies geschieht wie gesagt in der Absicht, homogene Gruppen von Aufteilungen herauszufinden. Getrennt nach diesen Gruppen werden nun die Aufteilungen zu Objekt x Objekt-Matrizen zusammengefasst und erst dann wird in der üblichen Weise, aber getrennt für jede dieser Gruppen,

durch weitere Clusteranalysen je ein HCS der Objekte erzeugt.

Diese Art der Datenanalyse ist anscheinend bislang von niemandem ins Auge gefasst, geschweige denn durchgeführt worden. Der Gedanke an eine differentielle Analyse von Sorting-Daten taucht in der Literatur überhaupt zum ersten Mal bei ARABIE und BOORMAN (1973) auf in Zusammenhang mit der praktischen Erprobung von zwölf Distanzmassen für Aufteilungen mit MDS-Techniken. Dass die Möglichkeit der inversen Clusteranalyse von Aufteilungen bislang übersehen werden konnte, ist eigentlich deshalb etwas verwunderlich, weil die Symmetrie der Analyse von Objekt/Prädikat-Tafeln,direkt (oder indirekt über Distanzmatrizen) Ausgangspunkt für viele Clusteranalysen, und die Analogie zur Q-Technik der Faktorenanalyse so offenkundig ist.

Leider gibt es jedoch nicht e i n Distanzmass für Aufteilungen, sondern deren mehrere. Und ihre Eignung für den oben beschriebenen Zweck ist fraglich. BOORMAN und ARABIE (1972), die wohl erstmalig die bis dahin verstreut veröffentlichten Distanzmasse für Aufteilungen in einer Zusammenschau systematisch dargestellt haben, kommen zu dem Resultat, dass von der Konzeption her keines als das a l l g e m e i n geeignetste gelten kann. Auch das Ergebnis ihrer anschliessenden empirischen Untersuchung erlaubt wegen des andersartigen Verwendungszusammenhanges keine tragfähigen Schlüsse hinsichtlich ihrer Brauchbarkeit im Rahmen einer inversen Clusteranalyse von Aufteilungen.

## 4.3 Die Distanz von Aufteilungen

Da der Beginn dieser Arbeit weiter zurückliegt als das Erscheinen von BOORMAN und ARABIEs vergleichender Studie über Distanzmasse für Aufteilungen, bestand zunächst die Notwendigkeit zur Entwicklung eines eigenen Masses, wenn das Problem der Inhomogenität auf die beabsichtigte Art in Angriff genommen und für die differentielle Betrachtungsweise fruchtbar gemacht werden sollte. Im Rückblick hat sich diese Schwierigkeit als positiv erwiesen, denn das Mass des Verfassers scheint - um das Ergebnis vorwegzunehmen - die verschiedenen Tests doch insgesamt etwas besser überstanden zu haben als die anderen im Vergleich befindlichen Masse.

Insgesamt sind nämlich schliesslich vier Distanzmasse auf ihre Eignung im Rahmen einer inversen Clusteranalyse für Aufteilungen untersucht worden. Drei davon stammen aus der Studie von BOORMAN und ARABIE, eins vom Verfasser. Die Auswahl der drei erfolgte unter mehreren Gesichtspunkten: Da ARABIE und BOORMAN gefunden hatten, dass sich die zwölf Distanzmasse aufgrund ihrer Ähnlichkeit in der Konzeption und im empirischen Verhalten plausibel in wenigen Gruppen zusammenfassen liessen, sollten diese Gruppen jeweils durch ein Mass repräsentiert sein. Daneben wurde mathematische Elaboriertheit und praktische Bewährung berücksichtigt. Die Wahl fiel auf PAIRBONDS, INFOTWO und INFOTHRY. Wir stellen sie vor nach der Definition der dazu nötigen Hilfsmittel.

Definition 8: Der Durchschnitt von zwei Aufteilungen $P_1 \sqcap P_2$ ist wie folgt definiert:

> Zwei Elemente a,b ∈ S sind in derselben Klasse $z_{ij}$ von $P_1 \sqcap P_2$ dann und nur dann, wenn sie in derselben Klasse $c_i$ von $P_1$ und auch in derselben Klasse $d_j$ von $P_2$ sind.

Beispiel: $P_1$ : $c_1 = (a,b,c)$; $c_2 = (d)$;

$P_2$ : $d_1 = (a,b)$; $d_2 = (c,d)$.

Die Klassen $z_{ij}$ sind die Zellen der folgenden Tabellen:

| | $d_1$ | $d_2$ |
|---|---|---|
| $c_1$ | (a,b) | (c) |
| $c_2$ | $\emptyset$ | (d) |

$P_1 \sqcap P_2$ ist in diesem Fall also {(a,b),(c),(d)}.

Definition 9: Die Definition der Vereinigung von zwei Aufteilungen ist etwas komplizierter:

> a,b ∈ S sind in derselben Klasse $z_{ij}$ von $P_1 \sqcup P_2$ dann und nur dann, wenn es eine endliche Folge $a = a_0, a_1, \ldots, a_q = b$ gibt, derart, dass für jedes j das Paar $a_j, a_{j+1}$ in derselben Klasse von $P_1$ oder $P_2$ ist.

Es lässt sich zeigen, dass $P_1 \sqcup P_2$ die feinste Klassifikation ist, die gröber als jede von ihnen ist ('feiner' und 'gröber' jeweils im Sinne von Definition 2). Und entsprechend gilt: $P_1 \sqcap P_2$ ist die gröbste Aufteilung, die feiner ist als $P_1$ und $P_2$.

Definition 10: $P_1 \sqcup P_2$ heisst im Englischen "least upper bound".

Definition 11: $P_1 \sqcap P_2$ heisst "least lower bound".

> Wenn in einer Menge $\mathfrak{P}(S)$ von Aufteilungen für jedes Paar $P_1$ und $P_2$ Durchschnitt $P_1 \sqcap P_2$ und Vereinigung $P_1 \sqcup P_2$ existieren, so bildet die Menge $\mathfrak{P}(S)$ hinsichtlich der Relation '$\sqsubseteq$' (Definition 2) einen Verband mit $P_0$ und $P_I$ als Null- bzw. Eins-Element (s. SZASZ, 1963).

Wie BOORMAN und ARABIE nachweisen, sind die meisten ihrer Distanzmasse $\delta(P_1,P_2)$ für Aufteilungen in Form der Gleichung

$$\delta(P_1,P_2) = g(P_1) + g(P_2) - 2g(P_1 \sqcap P_2) \tag{29}$$

darstellbar. Dabei ist g eine Funktion der Aufteilungen, die ihre "Höhe" im Verband zum Ausdruck bringt. Dazu muss man sich den Verband als ein zwischen den Polen $P_0$ (Süden) und $P_I$ (Norden) aufgespanntes Netzwerk denken. Die Höhe ist dann so etwas wie der Grad der Feinheit der Aufteilungen. Die Gleichung erinnert nicht zufällig an den Cosinus-Satz, bei dem die gesuchte Distanz die dritte Seite eines Dreiecks ist, von dem die beiden anderen und der durch sie gebildete Winkel gegeben sind.

Ein plausibles Mass für die Höhe g einer Aufteilung ist schlicht die Anzahl der Paarbindungen innerhalb ihrer Klassen. Ist z.B.

$$P = \{(a,b,c,d), (e), (f,g)\}$$

dann ist die Anzahl der Paarbindungen graphisch veranschaulicht:

e f — g also 6 + 0 + 1 = 7

und sie berechnen sich nach der Formel

$$\sum_i \binom{|c_i|}{2} = 1/2 \sum_i |c_i|(|c_i|-1) \tag{30}$$

Die Anzahl der Paarbindungen erreicht ihr Maximum für $P = P_I$ mit $\binom{n}{2}$, wobei n die Anzahl der Elemente von S ist. Wenn das Mass mit diesem Wert normiert wird, so dass es im Intervall $[0,1]$ liegt, ergibt sich als Distanz PAIRBONDS zwischen den Aufteilungen $P_1$ und $P_2$:

Definition 12:

$$D(P_1,P_2) = \frac{2}{n(n-1)} \sum_i \binom{|c_i|}{2} + \sum_j \binom{|d_j|}{2} - 2 \sum_{i,j} \binom{|z_{ij}|}{2}.$$

Denselben Ausdruck erhält man übrigens als normierte City-block-Metrik für Inzidenz-Matrizen von Aufteilungen:

Es sei

Definition 13:

$$M(P) = [a_{ij}]_{nxn} \text{ definiert mit } a_{ij} = \begin{cases} 0, & \text{für } a_i \text{ und } a_j \text{ in verschiedenen Klassen} \\ 1, & \text{für } a_i \text{ und } a_j \text{ in derselben Klasse} \end{cases}$$

Dann ist

Definition 14:

$$D(P_1,P_2) = \frac{1}{n(n-1)} \sum_{i,j} |a_{ij} - b_{ij}| \text{ für } M(P_1) = [a_{ij}] \text{ und } M(P_2) = [b_{ij}].$$

Den beiden nächsten Massen ist ihre informationstheoretische Grundlegung gemeinsam. Das konstitutive Konzept ist die 'Wahrscheinlichkeit $p_i$ der Zugehörigkeit zu einer bestimmten Klasse $c_i$.

Definition 15: $p_i = |c_i|/n$.

Diese Zuordnung von Wahrscheinlichkeiten zu Klassen ist eigentlich nicht ganz im Sinne der Logik der Klassifikation, weil es sich bei den Klassen ja nicht um a priori gegebene Schubfächer oder unabhängige Merkmale handelt.

Für die Entropie einer Aufteilung ergibt sich bei Verwendung von Definition 15 für $P_i$ wie üblich

$$(31) \qquad H(P) = -\sum_i p_i \operatorname{ld} p_i = -\sum_i |c_i|/n \operatorname{ld} |c_i|/n .$$

Die bedingte Entropie von P gegeben Q wird nun definiert als

Definition 16:

$$H(P|Q) = H(P \sqcap Q) - H(Q)$$

Das Distanzmass INFOTWO ist nun definiert durch die Unsicherheit, die verbleibt, wenn jeweils eine von beiden Aufteilungen gegeben ist hinsichtlich der anderen:

Definition 17:

$$C(P,Q) = H(P|Q) + H(Q|P).$$

C(P,Q) erhält eine der allgemeinen Gleichung entsprechende Form durch Einsetzen von Definition 8 in Definition 9 und Umformung nach

(32) $H = H_{max} - H = ld\ n - H$

nämlich

(33) $C(P,Q) = 2\ H(P \sqcap Q) - H(P) - H(Q)$

und folglich

(34) $C(P,Q) = (ld\ n - H(P)) + (ld\ n - H(Q)) - 2(ld\ n - H(P \sqcap Q))$.

Das zweite informationstheoretische Mass INFOTHRY ist eine Normalisierung von INFOTWO. Die Art der Normalisierung ist aber streng zu unterscheiden von der Normierung, die bei PAIRBONDS vorgenommen wurde und lediglich den numerischen Bereich, in dem sich die Werte bewegen sollten, festlegte. Die Normalisierung im Fall von INFOTHRY hat eine inhaltliche Funktion: sie soll das Mass weniger sensitiv für die absolute Höhe der Aufteilung machen. Es wird gleichsam die Dimension Höhe als unerwünscht "herausfaktorisiert". Deshalb besteht zwischen Distanzen von INFOTHRY und INFOTWO gewöhnlich auch keine monotone Beziehung. Die Normalisierung erfolgt durch Division mit $H(P \sqcup Q)$ und INFOTHRY ist demnach definiert als

Definition 18:

$$E(P,Q) = \frac{2\ H(P \sqcap Q) - H(P) - H(Q)}{H(P \sqcup Q)}$$

Was das Mass inhaltlich leistet, erhellt ebenfalls aus einer zweiten Herleitung von Definition 18. Ausgangspunkt ist hierbei die Transinformation von P und Q.

(35) $T(P;Q) = H(P) + H(Q) - H(P \sqcap Q)$

Durch einfaches Umformen ergibt sich E(P,Q) als Funktion der mit der Entropie ihres Durchschnitts relativierten Transinformation von P und Q:

$$(36) \qquad E(P,Q) = 1 - \frac{T(P;Q)}{H(P \sqcap Q)}$$

Dieses Distanzmass ist vor allem bekanntgeworden durch RAJSKI (1961) und stellt ein formales Analogon dar zu dem von GALANTER (1956) entwickelten Distanzmass für Mengen (siehe dazu die Arbeiten von RESTLE (1959, 1961)).

Wir können die drei Distanzmasse für Aufteilungen PAIRBONDS, INFOTWO, INFOTHRY als die prominentesten Vertreter ihrer Art hier natürlich nicht umfassend charakterisieren und kritisieren. Hingewiesen werden soll lediglich auf eine ihrer Eigenschaften, die im Hinblick auf die Zwecke dieser Arbeit äusserst ungünstig ist, weil sie grundlegenden Annahmen der Hierarchischen Klassifikation zuwiderläuft: alle drei gewichten Unterschiede zwischen zwei Aufteilungen umso stärker, je ähnlicher diese $P_0$ sind. Anders ausgedrückt, sind a, b ∈ S in $P_1$ in einer gemeinsamen Klasse und in $P_2$ in verschiedenen, so fällt die dadurch bedingte Distanz zwischen $P_1$ und $P_2$ umso grösser aus, je grösser die Klasse, in der sich a und b befunden hatten, gewesen ist. Sowohl unsere Intuition, wie auch die Prämissen der hierarchischen Clusteranalyse legen aber gerade eine umgekehrte Abhängigkeit nahe. So scheint es eher plausibel, dass zwei Elemente mehr Übereinstimmung aufweisen, also einander ähnlicher sind, wenn nur sie allein zusammengefasst werden, als wenn sie innerhalb einer grösseren Klasse sich zusammenfinden. In diesem Sinn sprechen auch BOORMAN und ARABIE, wenn sie vorschlagen:

> "it would be worth considering a rival distance model in which the relation between distence produced and the cardinality of the original cell is inverse - the argument being that a single unit can be more readily lost from a large collection (i.e. cell) than from a small one".

Das von uns entwickelte Mass hat nun in der Tat die wünschenswerte Eigenschaft, dass Distanzen umso grösser ausfallen, je kleiner die Klassen sind, auf die sich die Unterschiede stützen. Überhaupt wurde mehr Wert auf Übereinstimmung mit intuitiven Erwartungen an ein Mass für Aufteilungen gelegt, als auf die Herleitbarkeit aus einer mathematischen Theorie (wie z.B. der Informationstheorie). Dementsprechend lässt sich der Charakter des Handgestrickten nicht ganz verleugnen.

Im Zusammenhang mit dem Distanzmass PAIRBONDS wurde bereits die Darstellung einer Aufteilung durch eine Inzidenz-Matrix vorgestellt. Für das Mass des Verfassers verwenden wir diese Matrix $M(P) = [a_{ij}]_{nxn}$. Hinzu kommt eine zweite Matrix $M'(P)$, die als eine Art Gewichtung fungiert. Ihre Zellenwerte $W_{ij}$ sind eine Funktion von Anzahl und Grösse der Klassen. Die weitere Verrechnung erfolgt ähnlich wie bei der City-block-Metrik für Inzidenz-Matrizen, nur dass bei unserem Mass diejenigen $w_{ij}$ addiert werden, für die hinsichtlich der Matrizen M(P) Übereinstimmung vorliegt.

Ein Beispiel:

$$P = \{(a,b,c)_1\ (d,e)_2\ (f)_3\}$$

$M(P) =$

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | | | | | |
| b | 1 | | | | | |
| c | 1 | 1 | | | | |
| d | 0 | 0 | 0 | | | |
| e | 0 | 0 | 0 | 1 | | |
| f | 0 | 0 | 0 | 0 | 0 | |

Der Symmetrie wegen wird jeweils nur die halbe Matrix berücksichtigt.

$$Q = \{(a,b)_1\ (c,d)_2\ (e,f)_3\}$$

$M(Q) =$

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | | | | | |
| b | 1 | | | | | |
| c | 0 | 0 | | | | |
| d | 0 | 0 | 1 | | | |
| e | 0 | 0 | 0 | 0 | | |
| f | 0 | 0 | 0 | 0 | 1 | |

Bei der Definition von M'(P) dienen wiederum senkrechte Balken "| |" jeweils zur Bezeichnung der Kardinalzahl. So bedeutet

Definition 19:

$|P| = m$ die Anzahl der Klassen von P,

Definition 20:

$|c_i|$ die Anzahl der Elemente in der Klasse $c_i$, und

Definition 21: $|P^*|$ die Anzahl der Klassen mit $|c| \geq 2$. Entsprechend sind

Definition 22: $|c_i^*|$ Klassen mit $|c_i| \geq 2$.

Die Gewichtung erfolgt nach unterschiedlichen Prinzipien, je nachdem ob $a_{ij} = 1$ oder $a_{ij} = 0$ ist. Aber die Summe aller $w_{ij}$ für den einen wie den anderen Fall ist jeweils 1:

$$(37) \qquad \sum_{i,j(a_{ij}=1)} w_{ij} = 1 \text{ und } \sum_{i,j(a_{ij}=0)} w_{ij} = 1.$$

Für jede Klasse $c_k$ wird auf alle $w_{ij}$, $i,j \in c_k$ der gleiche Anteil von $1/|P^*|$ verteilt. Da die Anzahl der Paare $i,j \in c_k$ gerade $\binom{|c^*|}{2}$ ist, hat $w_{ij}$ den Wert $1/|P^*|\binom{|c^*|}{2}$

Im Falle $a_{ij} = 0$ haben alle $w_{ij}$ den gleichen Wert. Damit Bedingung (37) erfüllt ist, ist dieser Wert

Definition 23:

$$w_{ij} = \frac{1}{\binom{n}{2} - \sum_{c_k^* \in P} \binom{|c^*|}{2}}$$

In unserem numerischen Beispiel ergeben sich somit für

M'(P)

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | | | | | |
| b | 1/6 | | | | | |
| c | 1/6 | 1/6 | | | | |
| d | 1/11 | 1/11 | 1/11 | | | |
| e | 1/11 | 1/11 | 1/11 | 1/2 | | |
| f | 1/11 | 1/11 | 1/11 | 1/11 | | |

und

$M'(Q)$

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | | | | | | |
| b | | 1/3 | | | | |
| c | | 1/12 | 1/12 | | | |
| d | | 1/12 | 1/12 | 1/3 | | |
| e | | 1/12 | 1/12 | 1/12 | 1/12 | |
| f | | 1/12 | 1/12 | 1/12 | 1/12 | 1/3 |

Die Summe der $w_{ij}^P + w_{ij}^Q$ mit $a_{ij}^P = a_{ij}^Q$ erreicht ihr Maximum 4 wenn und nur wenn P = Q ist. Wir definieren deshalb als Distanz von P und Q

Definition 24:

$$R(P,Q) = 4 - \sum_{i,j(a_{ij}^P = a_{ij}^Q)} (w_{ij}^P + w_{ij}^Q)$$

und erhalten jetzt $r(P,Q) = 4 - (1 - 1/6 + 9(1/11 + 1/12)) \approx 1{,}36$.

## 5. DER VERGLEICH DER VIER DISTANZ-MASSE FÜR AUFTEILUNGEN

Grundsätzlich gibt es wohl drei Möglichkeiten, die Eignung eines methodischen Instruments, wie es die Distanzmasse sind, zu begründen. Eine davon ist die abstrakt-theoretische Legitimation. Wir haben davon insofern Gebrauch gemacht, als wir bei der Einführung der Masse auf die Bedeutung der sie definierenden Grössen hingewiesen haben:
bei PAIRBONDS war dies der jeweilige "Höhenunterschied" der Aufteilungen bezüglich ihrer gröbsten gemeinsamen Verfeinerung, bei INFOTWO ihre bedingte Entropie, bei INFOTHRY die Transinformation und bei dem von uns entwickelten Mass die mit Anzahl und Grösse der Klassen gewichtete Übereinstimmung.

Eine andere Möglichkeit besteht darin, einfache Beispiele zu konstruieren, die einerseits [illegible] sind, andererseits geeignet sind, kritische Unterschiede im Verhalten der Methoden zu erhellen, so dass sie anhand intuitiver oder sonstwie vorgegebener Kriterien beurteilt werden können. Uns kam es darauf an, unseren durch BOORMAN und ARABIE in seinem Gewicht bestätigten Einwand (s. Zitat, S. 62 ) gegenüber den drei ersten Massen zu belegen, dass sie das gemeinsame Auftreten von zwei Objekten in derselben Klasse nicht umso höher bewerteten, je kleiner diese ist, sondern umgekehrt.

Zu diesem Zweck wurden drei Aufteilungen ausgewählt, alle vom selben Typ, mit drei, zwei und einem Element pro Klasse:

$$O = \{(a,b,c)\ ,\ (d)\ ,\ (e,f)\},$$

$$P = \{(a,b)\ ,\ (c,d,e)\ ,\ (f)\},$$

$$Q = \{(a,b,c)\ .\ (d,e)\ ,\ (f)\},$$

O und Q unterscheiden sich in Bezug auf das Element e. Und zwar ist e aus einer kleinen in eine andere kleine Klasse übergewechselt. Im Gegensatz dazu ist im Fall von P und Q das Element c von einer (relativ) grossen in eine andere grosse Klasse gekommen. Sonst gibt es keine Unterschiede zwischen den beiden Paaren OQ und PQ.

Wie bilden nun die vier Distanzmasse diesen Sachverhalt ab ? Im Sinne unserer Forderung sollte

$$D(O,Q) > D(P,Q)$$

sein. Tatsächlich ergeben sich jedoch

für PAIRBONDS $D(O,Q) = .13 < D(P,Q = .26$ und

für INFOTWO $D(O,Q) = .62 < D(PQ) = .90$ und

für INFOTHRY $D(O,Q) = .11 < D(P,Q) = 1.1.$

Nur unser Mass zeigt mit

$$D(O,Q) = 2.2 > D(P,Q) = 1.2$$

die gewünschte Relation.

Welche Bedeutung kann diesem Vergleich nun beigemessen werden ?

Man sollte ihn wohl im Sinn der Logik der Falsifikation verstehen. Mit dem Unterschied, dass nach ihrer klassischen Version besonders riskante Realisationen der Möglichkeit des Scheiterns ausgesetzt werden, während unser Beispiel ausgewählt wurde um eine wichtige Eigentümlichkeit zu verdeutlichen. Freilich zeigen sich auch hier die Probleme des methodologischen Fallibilismus: da die Herstellung von Situationen, an denen ein Distanzmass scheitern kann, letztlich nicht einem verbindlichen Kanon folgen kann, sondern der Erfindungsgabe und damit der wissenschaftlichen Entwicklung unterliegt, kann solch Scheitern nie endgültig sein, zumal das Gewicht einer

Testsituation sich im Licht neuer Erkenntnis jederzeit ändern kann. So wird man natürlich die Frage stellen, ob nicht die Verwendung gerade dieser drei Aufteilungen eine ungerechtfertigte Begünstigung bedeutet. Diesem Einwand ist schwer zu begegnen. Als gesichert darf aber wohl gelten, dass drei der Masse diesem einen im Zusammenhang mit HCSs sehr wichtigen Kriterium nicht genügen.

Der dritte und eigentlich direkteste Test der Distanzmasse ist ihre praktische Erprobung im - soweit das möglich ist - natürlichen Verwendungszusammenhang. Dem Vergleich der Masse am Beispiel mehr oder weniger realistischer Daten ist das folgende Kapitel gewidmet.

## 5.1 Die empirische Erprobung der Distanzmasse

Im Abschnitt 4.2 haben wir dargelegt, welche Funktion ein Distanzmass für Aufteilungen im Rahmen einer inversen Clusteranalyse zu erfüllen hat: es ist die Grundlage zur Identifizierung von Teilgruppen von Aufteilungen, die eine hinreichend grosse strukturelle Homogenität aufweisen, um eine Hierarchische Clusteranalyse der sortierten Objekte gerechtfertigt erscheinen zu lassen.

Bei einer empirischen Erhebung mit dem Sortierverfahren gibt es natürlich in der Regel keine Vorinformation über Anzahl und Grösse solcher homogenen Versuchspersonengruppen; die sollen ja gerade herausgefunden werden. Zum Nachweis der prinzipiellen Brauchbarkeit dieser Verfahrenskombination benötigen wir jedoch gerade solches Vorwissen als Validierungsgrundlage. Ausgangspunkt ist eine Gruppe von Vpn bzw. ihre Aufteilungen, von der bekannt ist, dass sie in (wir nehmen den einfachsten Fall) zwei gleich grosse Teilmengen von strukturell ähnlichen Aufteilungen zerfällt. Am Beispiel solcher Daten muss sich dann erweisen, ob und wie gut die mechanische Anwendung unseres Ansatzes zur Isolierung der homogenen Untergruppen führt, d.h. die im Vorwege erzeugte

Datenstruktur zu rekonstruieren erlaubt.

Es ist zu erwarten, dass die Verfahrenskombinationen aus Distanzmass und hierarchischer Clustermethode diese Aufgabe unterschiedlich erfolgreich bewältigen. Zeichnet sich ab, dass einige von ihnen bei verschiedenen Datensätzen - künstlichen und natürlichen - durchgängig überzeugende Ergebnisse liefern, so dürfen wir darauf vertrauen, dass auch, wenn wir ihrer Zusammensetzung nach unbekannte Mengen von Aufteilungen mit ihrer Hilfe analysieren, den gefundenen Gruppierungen wahre Unterschiede entsprechen.

Was wir hier planen, folgt natürlich der Logik des Methodenexperiments, genauer gesagt, der der Plasmodenstudie, wie sie von CATTELI definiert wird:

> "A set of natural variables, concretely measured, the formal relations of which are reliably known and which typify those posited to exist in the phenomena currently being investigated for the mathematical model." (CATTELL and SULLIVAN, 1962).

Durch ein Methodenexperiment soll also nicht wie sonst bei Experimenten Art und Ausmass der Wirkung einer unabhängigen Variablen ermittelt d.h. eine Entscheidung über das Vorliegen eines bis dahin nicht hinreichend gesicherten Zusammenhangs gefällt werden. Im Gegensatz dazu setzt das Methodenexperiment die Kenntnis eines bestimmten Zusammenhangs voraus, um darauf aufbauend die Frage stellen zu können, wie gut dieser Zusammenhang durch eine neue Methode abgebildet wird.

Anders ausgedrückt: Um die Methode zu beurteilen, wird ihr Ergebnis mit dem Vorwissen über den Gegenstand, das als eine Art Aussenkriterium dient, verglichen. So wird man - um ein anschauliches Beispiel zu geben - nach der Erfindung des Bi-Metalls ein daraus konstruiertes Thermometer unter verschiedenen Temperatur- (und anderen) Bedingungen mit dem normalen Quecksilberthermometer vergleichen, um seine Brauch-

barkeit als Temperaturanzeiger beurteilen zu können. Aus der Psychologie sind die Untersuchungen zur Faktorenanalyse von THURSTONE (1957) bekannt (siehe auch ÜBERLA 1968), die zeigen sollen, dass diese Technik in der Lage ist, aus den korrelativen Zusammenhängen einer grösseren Anzahl von Variablen die - im voraus bekannten - Grundvariablen zu erschliessen.

## 5.2 Der Versuchsplan

Es wurden drei Untersuchungen durchgeführt. Der Aufbau des Versuchsplanes war immer derselbe. Zunächst wurde jeweils eine Menge von Aufteilungen erzeugt, bzw. erhoben, wobei durch die Art ihrer Entstehung eine Zweiteilung in homogene Untergruppen angestrebt wurde. Variiert wurde der Grad der strukturellen Homogenität. Über der Gesamtmenge der Aufteilungen wurden dann nach PAIRBONDS, INFOTWO, INFOTHRY und dem Mass des Verfassers (im folgenden 'ROSSBACH' genannt) die paarweisen Distanzwerte berechnet. Die sich so ergebenden um die Diagonale symmetrischen Distanzmatrizen bildeten die Eingabeinformationen für die fünf geläufigen und in ihrer Clusterdefinition jeweils deutlich verschiedenen hierarchischen Clusteranalysen Minimum-, Maximum-, Gruppendurchschnitt-, Schwerpunkt- und WARD-Methode. Die durch sie erzeugten $4 \cdot 5 = 20$ Baumgraphen wurden hinsichtlich ihrer 'Diskrepanz' zur theoretischen Zeiteilung verglichen.

## 5.3 Erster Test

Ein geeignetes Mass für die Diskrepanz zwischen Aufteilungen und "ranked trees", also Baumgraphen, deren Gabelungspunkte nach ihrer Höhe ranggeordnet sind, stammt von HUBERT (1972). Es basiert auf dem Vergleich aller Paare von Objektepaaren, von denen die beiden Objekte des ersten Paares sich in verschiedenen Klassen der Aufteilung befinden und die anderen beiden in derselben, mit den diesen Objektepaaren

entsprechenden Distanzen der ultrametrischen Matrix, die den Baumgraphen repräsentiert. 'Kompatibilität' zwischen Baum und Aufteilung liegt vor, wenn die zweite Distanz (also die der zusammengefassten Objekte) kleiner als die erste ist. Der Diskrepanzwert ist die Anzahl der Verletzungen dieser Kompatibilität anzeigenden Relation. Inhaltlich bedeutet das einfach, dass wenn zwei Objekte in einer Aufteilung durch Plazierung in dieselbe Klasse als einander ähnlicher ausgewiesen wurden als zwei getrennte, dies auch in der Baumstruktur (d.h. den Distanzen) seine Entsprechung finden muss und umgekehrt. (Siehe den folgenden Abschnitt)

In einer geringen Diskrepanz zwischen theoretischer Zweiteilung und dem Baumgraph der Aufteilungen, ist ein Beweis für die Eignung der betreffenden Verfahrenskombination zu sehen, diejenigen Aufteilungen herauszufinden und zu einer Gruppe zusammenzufassen, die auf derselben hierarchischen Struktur basieren.

## 5.4 Die Diskrepanz zwischen einem HCS und einer Aufteilung (nach HUBERT, 1972)

S sei eine endliche Menge von Objekten der Anzahl n, $d_{ij}$ ein Distanzmass definiert auf S x S, P eine Aufteilung von S.

Das HUBERTsche Diskrepanzmass ist nun definiert als die Mächtigkeit des Durchschnitts von zwei Mengen. Die eine der geschnittenen Mengen, $Q$, besteht aus den geordneten Paaren der ungeordneten Objektpaare, bei denen die erste Distanz des HCS kleiner als die zweite ist.

Definition 25: $Q = \{((i,j), (k,l)) \mid i,j,k,l \in S;$

$$i \neq j,\ k \neq l;\ d_{ij} < d_{kl}\}$$

Die andere Menge $R_p$ umfasst diejenigen geordneten Paare aus Elementen von S x S, bei denen die Objekte des ersten

Paares nicht derselben Klasse von P angehören, während die des zweiten Paares dies tun.

Definition 26:

$$R_p = \{((i,j)\ ,\ (k,l)) \mid i,j,k,l \in S;\ i \neq j,\ k \neq l;\ k,l \in c_s,\ i \notin c_t \text{ mit } r \neq t\}$$

Die Mächtigkeit der Schnittmenge $Q \sqcap R_p$ kann nun noch relativiert werden durch den grösstmöglichen Wert, den sie unter Berücksichtigung der mit Anzahl m und Grösse $|c_i|$ der Klassen von P gegebenen Restriktionen annehmen kann. Also

Definition 27:

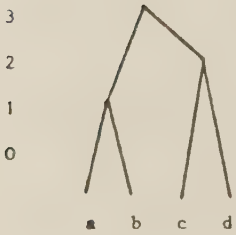
$$\text{Diskrepanz}(HCS,P) = \frac{|Q \sqcap R_p|}{\max |Q \sqcup R_p|} .$$

Der Wert des Maximums von $|Q \sqcap R_p|$ errechnet sich wie folgt:

$$(38)\quad n(n-1)/2 - \sum_{i=1}^{m} |c_i|(|c_i| - 1)/2 \quad \sum_{i=1}^{m} |c_i|\ (|c_i| - 1)/2 .$$

Ein Beispiel:

Gegeben sind ein HCS und eine Aufteilung P (siehe Abb. 4 unten):




| | a | b | c | d |
|---|---|---|---|---|
| a | | 1 | 3 | 3 |
| b | | | 3 | 3 |
| c | | | | 2 |
| d | | | | |

$$P = \{(a,b,c),(d)\}$$

Abb. 4: Darstellung eines HCS (durch Baumgraph und Distanzmatrix) und einer Aufteilung P derselben Objektmenge.

Es wird nun die Menge der ungeordneten Objektpaare S x S gebildet:

$$S \times S = \{(a,b)\ ,\ (a,c)\ ,\ (a,d)\ .\ (b,c)\ ,\ (,d)\ ,\ (c,d)\}$$

Die nach Definition 25 geordneten Paare dieser Menge sind in der Matrix (Abb. 5 unten) durch einen Schrägstrich "/" repräsentiert. Dies ist eine weitere isomorphe Abbildung des HCS. Dasselbe gilt für die Abbildung der Aufteilung durch den gegenteiligen Schrägstrich "\" bei den Paaren aus S x S, die Definition 26 genügen.

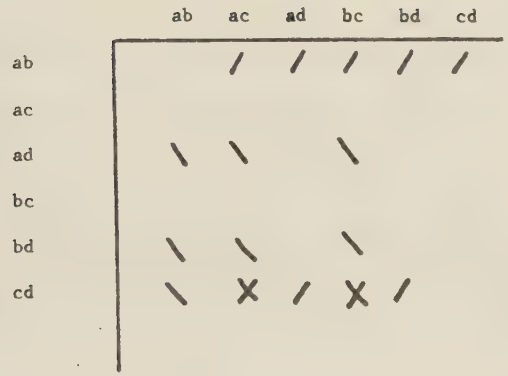



Abb. 5: Matrix aller Paare von {a,b,c,d}. BEzüglich derjenigen Paare von Paaren, die mit einem X versehen sind, besteht eine "Diskrepanz" zwischen dem HCS und der Aufteilung P der vorigen Abbildung.

Die Menge der $Q \cap R_p$ wird von den Paaren gebildet, die durch ein Kreuz "X" markiert sind. Ihre Anzahl ist in diesem Fall zwei. Das Maximum ist ebenfalls aus der Matrix ersichtlich, nämlich neun, bzw. errechnet sich nach (3) wie folgt:

$$\max \; Q \cap R_p = \left[6-(3+0)\right]\left[3+0\right] = 9$$

Als Wert für die Diskrepanz zwischen diesem HCS und der Aufteilung P erhält man also:Diskrepanz (HCS,P) = 2/9 = 0.22.

## 5.5 Zweiter Test

Der erste Test setzte die "Richtigkeit" oder Validität der theoretischen Dichotomie in zwei homogene Gruppen A und B voraus. Diese Unterstellung ist jedoch fragwürdig und das aus verschiedenen Gründen, je nach Entstehung der Aufteilungen. Wenn sie, wie im Fall der ersten beiden Datensätze (s.u.) durch einen Zufallsprozess erzeugt wurden, kann - eben zufällig und unbemerkterweise - unter der A-Gruppe eine Aufteilung entstanden sein, die tatsächlich einer oder sogar mehreren Aufteilungen der B-Gruppe so ähnlich ist, dass es nicht sinnvoll wäre, ihre Zuordnung zur B-Gruppe einer Verfahrenskombination als Fehler anzulasten. Ganz im Gegenteil. Entsprechendes gilt, wenn Aufteilungen von Vpn gebildet werden: Eine Vp, obwohl der A-Gruppe angehörend (d.h. so selegiert) liefert eine Aufteilung, die eher für B-Vpn typisch ist. Auch hier sollte eigentlich ihre Subsumption unter B nicht als Fehlklassifikation bewertet werden.

Diese Überlegungen führen auf eine differenziertere Methode zur Bewertung der Distanzmasse. Statt schlichtweg die theoretische Zweiteilung als Kriterium zu nehmen, betrachten wir die tatsächlich erreichte strukturelle Homogenität der beiden empirischen Teilgruppen, wie sie uns die jeweils höchste Gabelung der Bäume liefert. Dabei orientieren wir uns nur noch insofern an der theoretischen Dichotomie, als wir eine Unterteilung in mehr als zwei Gruppen nicht als Lösung akzeptieren, weil trivialerweise mit wachsendem k (der k-Teilung) auch die Homogenität zunimmt. (Bei k = n würde schliesslich jede Aufteilung mit sich selbst verglichen und es bestünde gar keine Möglichkeit zur Diskrepanz).

Gemessen wird die strukturelle Homogenität der beiden Teilgruppen folgendermassen: Zuerst wird für jede von ihnen die Distanzmatrix der sortierten Objekte (nach Definition 8) berechnet und daraus durch Anwendung einer hierarchischen Clustermethode ihr HCS. (Unsere Entscheidung dafür gerade die Maximum-

Methoden einzusetzen, war nicht willkürlich; man vergleiche die Charakteristiken der verschiedenen Clusteranalysen S.32 ff.). Wenn die Aufteilungen der beiden Untermengen tatsächlich vollständig homogen sind, müssten sie identisch sein mit den Aufteilungen, die als Schnitte durch den Baum entstehen, dem sie zugrundeliegen. Demnach zeigt die Summe ihrer Diskrepanzen zum HCS (nach HUBERT) das Ausmass ihrer Inhomogenität an. Die Diskrepanzwertsummen beider Teilmengen addiert, sind das Kriterium des zweiten Tests.

Schema des Versuchsplanes

Gewinnung eines Datensatzes:

1 Durch einen Zufallsprozess wird eine Menge von Aufteilungen erzeugt, die in zwei gleichgrosse, Strukturell homogene Untergruppen zerfällt.

1° Durch zwei nach ihrer Einstellung selegierte gleichstarke Extremgruppen von Vpn, die dieselbe Menge von für diesen Bereich bedeutsamen Begriffen sortieren.

2 Berechnung der Distanzmatrix zwischen sämtlichen Aufteilungen eines Datensatzes nach PAIRBONDS, INFOTWO, INFOTHRY und ROSSBACH

3 Hierarchische Clusteranalyse nach Minimum-, Maximum-, Gruppendurchschnitt-, Schwerpunkt- und WARD-Methode

4 Berechnung der Diskrepanz (nach HUBERT) zwischen theoretischer Zweiteilung und dem HCS der Aufteilungen als erster Test für den empirischen Erfolg der betreffenden Methodenkombination

4° Bestimmung der empirischen Zweiteilung der Sorting-Menge anhand der obersten Gabelung des Baumgraphen der Aufteilungen

5° Berechnung der strukturellen Homogenität der beiden empir. Teilmengen durch Vergleich ihrer nach der Maximum-Methode erzeugten beiden Objekt-Bäume mit den dafür zugrundegelegten Aufteilungen nach HUBERTs Diskrepanzmass

## 6. DIE ERGEBNISSE DES EMPIRISCHEN VERGLEICHES

Besondere Probleme bei der Darstellung der empirischen Ergebnisse resultieren aus dem Umfang und der Komplexität des zu beurteilenden Materials. Für jeden von drei Datensätzen sind die Leistungen von zwanzig Verfahrenskombinationen miteinander zu vergleichen. Jede einzelne dieser Kombinationen aus einem von vier Distanzmassen für Aufteilungen und einer von fünf Hierarchischen Klassifikationsverfahren liefert einen Baumgraphen mit den Aufteilungen (den Ausgangsdaten) als Endknoten. Dieser empirische Baumgraph soll die mit den besonderen Bedingungen des Methodenexperimentes gesetzte - theoretische - Zweiteilung der Menge der Aufteilungen möglichst gut widerspiegeln. Im Idealfall bedeutet dies, dass der Baum genau zwei Hauptäste besitzt und alle Aufteilungen der einen Gruppe unter den einen und alle Aufteilungen der anderen Gruppe unter den anderen Ast fallen. Eine Verfahrenskombination ist also erfolgreich, wenn sie die mit der experimentellen Bedingung induzierte Separation in zwei homogene Teilgruppen exakt reproduziert.

Nun ist der Fall der absolut korrekten Wiedergabe der ursprünglichen Zweiergruppierung zum einen äusserst selten, zum anderen die Möglichkeit der Differenzierung zwischen den in unterschiedlichem Ausmass fehlerhaften Lösungen natürlich wünschenswert. In den Abschnitten über den 'Ersten Test' und den 'Zweiten Test' wurden deshalb zwei Arten der Quantifizierung der 'Fehlerhaftigkeit' entwickelt - also, wie man wohl sagen kann, Indizes des goodness-of-fit. Damit erübrigt es sich, jeden einzelnen Baumgraphen vorzustellen. Ihre intuitive Bewertung zum Zweck des Vergleichs ist wegen ihrer Grösse und ihrer Anzahl ohnehin problematisch. Lediglich zur Illustration werden deshalb inhaltlich irgendwie bedeutsame Bäume in einzelnen Fällen vorgestellt. Eine wichtige Ausnahme von dieser Regel bildet der 1. Datensatz. Seine Entstehung und Auswertung soll in allen Schritten ausführlich belegt werden. Er ist als einziger von den dreien genügend überschaubar, weil

a) die Anzahl der klassifizierten Objekte vergleichsweise gering ist,

b) die Anzahl der Aufteilungen ebenfalls klein ist,

c) die beiden Teilgruppen von Aufteilungen strukturell vollständig homogen sind.

Vor allem Bedingung c) ist eine gute Vorausetzung, um sich die konkrete Bedeutung der beiden goodness-of-fit-Tests zu vergegenwärtigen, weil die darin formulierte Eigenschaft die Gegenüberstellung von theoretischer Zweiteilung und empirischem Baum (1. Test) und ebenso die Beurteilung des quantitativen Werts für die strukturelle Inhomogenität seiner beiden Hauptäste erleichtert.

Im übrigen wäre es wohl auch deshalb nachteilig, sich ausschliesslich auf die tabellarische Darstellung der Ergebnisse als Listen von goodness-of-fit-Indizes zu beschränken, weil die Nachvollziehbarkeit der Datenverarbeitung dadurch erheblich erschwert würde. Das bedeutet zugleich, dass die ausführliche exemplarische Darstellung von Datengewinnung und -auswertung anhand des ersten Datensatzes rückwirkend die Logik des Versuchsplanes erhellen soll.

## 6.1 Der erste Datensatz

Der erste Datensatz stellt so etwas wie eine "reine" Plasmode dar. "rein" bedeutet, dass das Ideal vollständiger Kompatibilität aller Aufteilungen derselben Gruppe untereinander vollständig verwirklicht ist. Definitionsgemäss bilden daher die Aufteilungen jeder Gruppe ein Hierarchisches Cluster System (siehe Abb. 6). Anders ausgedrückt: Es fehlt die sonst mit empirischen Daten nahezu unvermeidlich gegebene Streuung, die keinem systematischen Faktor zugeschrieben werden kann und deren Ergebnis "zufällige" Abweichungen sind.

Für Daten der hier diskutierten Art (Aufteilungen ) wäre ein gewisses Ausmass an Fehlerstreuung gleichbedeutend damit, dass alle Aufteilungen einer Gruppe zwar Ähnlichkeit zum selben HCS aufweisen, aber in unsystematischer Weise jeweils von ihm abweichen. Dieser Fall ist im 2. Datensatz realisiert.
Die Bedeutung dieses ersten Datensatzes liegt zum einen - wie gesagt - in seinem Wert als Demonstrationsbeispiel und zum

anderen in seiner Brauchbarkeit zur grundlegenden Prüfung des Ansatzes. Letzteres beruht vor allem auf folgender Überlegung: Wenn eine Kombination aus Distanzmass für Aufteilungen und Clusteranalyse überhaupt in der Lage ist, homogene Gruppen von Aufteilungen zu entdecken, dann muss das noch am ehesten gelingen, wenn keine störenden Abweichungen von vollständiger Kompatibilität bei den Elementen der jeweiligen Gruppen vorliegt. Umgekehrt lässt sich sagen, dass ein Verfahren, das selbst unter diesen günstigen Bedingungen versagt, erst recht nicht mit weniger klar strukturierten Daten befriedigende Gruppierungsleistungen erbringen wird.

### 6.1.1. Die Entstehung des 1. Datensatzes

Bei den Daten handelt es sich um 14 Aufteilungen einer zwölfelementigen Menge. Die Aufteilungen $P_1$ - $P_7$ und $P_8$ - $P_{14}$ bilden jeweils eine strukturell homogene Gruppe. Sie entstanden als Schnitte durch je einen Baumgraphen, der wiederum folgender mechanischer Prozedur entstammt:

(Siehe Aufteilung des 1. Datensatzes - S. 82)

Die Aufteilung des 1. Datensatzes

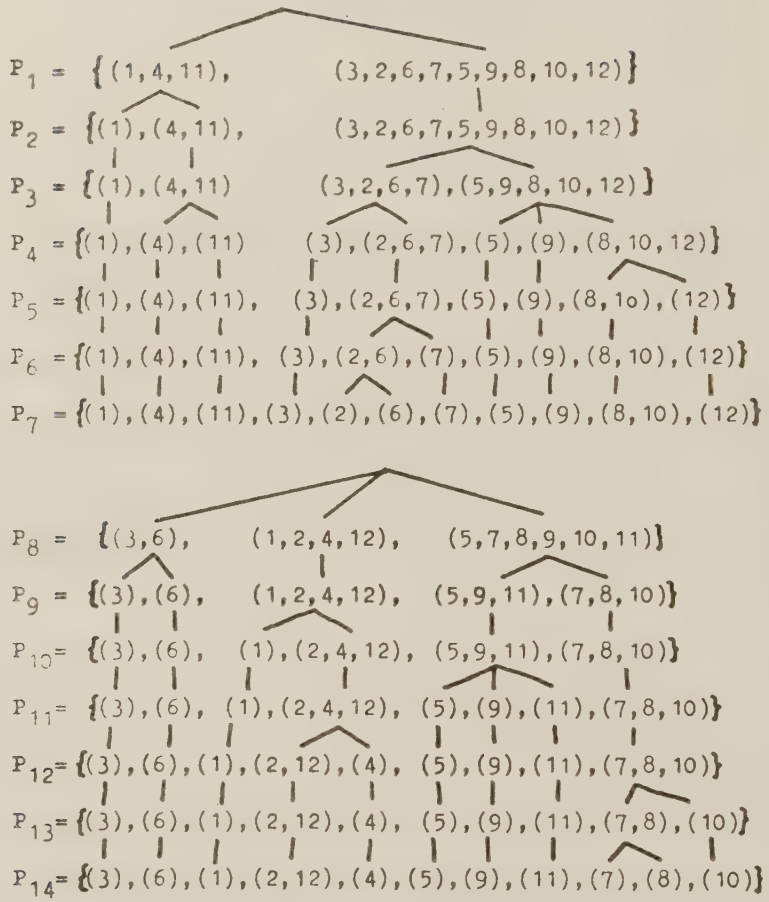



Abb. 6: Die 14 Aufteilungen des 1. Datensatzes bestehen aus zwei strukturell homogenen Untergruppen, nämlich $P_1$-$P_7$ und $P_8$-$P_{14}$. Jede Gruppe ist deshalb als Ergebnis von Schnitten durch denselben Baumgraphen der Objekte darstellbar.

In rekursiven Schritten wurde jeweils durch Zufall bestimmt:

1. Wieviele der auf der betreffenden Stufe vorliegenden Äste sich auf dieser Stufe überhaupt gabeln sollten.

2. Welche von diesen Ästen auf dieser Stufe geteilt werden sollten.

3. Auf wieviele Äste sich ein bestimmter Ast verzweigen sollte.

4. Wieviele Elemente unter einen neuentstandenen Ast fallen sollten.

5. Welche der unter einem Hauptast befindlichen Objekte seinen neuen Teilästen zugeordnet werden sollten.

Nach diesem Schema wurden die beiden Baumgraphen der Abbildung von oben, d.h. vom Scheitelpunkt oder Stamm her erzeugt. Die beiden Bäume sind, wie das zu erwarten- und im Sinne der Diskriminierbarkeit ihrer Aufteilungen auch zu wünschen - war, tatsächlich sehr verschieden ausgefallen. Die Produkt-Moment-Korrelation zwischen ihren Distanzmatrizen beläuft sich auf r = .06.

### 6.1.2 Die Analyse des ersten Datensatzes

Die Aufteilungen Nr. 1 bis 14 von Abbildung 6 bilden den ersten Datensatz und als solche die Eingabeinformation zur weiteren Analyse nach dem Schema des Versuchsplanes von Seite 78. Zunächst werden die Distanzen zwischen sämtlichen 14 Aufteilungen berechnet (Schritt 2 des Versuchsplanes). Die Berechnung der Distanzen erfolgt nach PAIRBONDS, ROSSBACH, INFOTWO und INFOTHRY, sodass damit vier Distanzmatrizen der Aufteilungen vorliegen (siehe Anhang, Tab. B - E)

Die Aufteilungen werden auf der Grundlage jeder dieser Distanzmatrizen durch fünf Hierarchische Clusteranalysen klassifiziert. Das Ergebnis sind zwanzig-zumeist verschiedene - Baumgraphen. Auf den folgenden Seiten finden sich jeweils zwei

Baumgraphen stellvertretend für die fünf, die sich für eine Distanzmatrix ergeben. Diese Einsparung war möglich, weil sich unter dem Baumgraphen hinreichend viele identische bzw. sehr ähnliche befanden (siehe dazu die tabellarische Übersicht ihrer quantitativen Bewertung, aus der dies ersichtlich wird). Die ausgewählten Beispiele repräsentieren also fast vollständig die Gesamtmenge der zwanzig Bäume des ersten Datensatzes.

Erläuterung zu den Abbildungen 7 - 11 der folgenden Seiten:

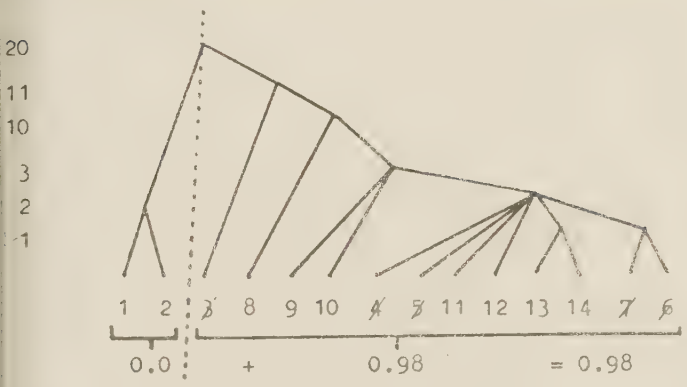
Auf der Grundlage der vier Distanzmatrizen der 14 Aufteilungen, wie sie sich durch Anwendung der vier Distanzmasse (PAIRBONDS, ROSSBACH, INFOTWO, INFOTHRY) ergeben hatten, wurden fünf Baumgraphen dieser Aufteilungen nach den Hierarchischen Klassifikationsverfahren (Minimum-, Maximum-, Gruppendurchschnitt-, Schwerpunkt-, WARD-Methode) erzeugt. Jeweils zwei davon für jedes Distanzmass werden auf den folgenden Seiten vorgestellt.

Am linken Rand neben den Bäumen findet sich der Wert des Clusters, dessen Knoten auf der betreffenden Höhe liegt. Eine senkrechte gepunktete Linie teilt den Baum im Scheitel in seine beiden Hauptäste. Die Nummern der unter denselben Hauptast fallenden Aufteilungen sind durch eine waagerechte Strecke unterhalb der Ziffern "eingeklammert". Darunter stehen die Werte für die Inhomogenität der so zusammengefassten Gruppen von Aufteilungen und ihre Summe als Gesamtmass für die Güte des Baumes (zur Berechnung siehe Seite 76). Rechts davon steht der Wert für die Diskrepanz zwischen dem Baum und der theoretisch erwarteten Zweiteilung der Menge der Aufteilungen, der goodness-of-fit (abgekürzt g.-o.-f.). Die Nummern der Aufteilungen, die im Sinne der theoretischen Erwartung durch den Baum klassifiziert werden, sind schräg durchstrichen.

Erster Datensatz

PAIRBONDS

Minimum-Methode




goodness-of-fit: 750

Wards Methode

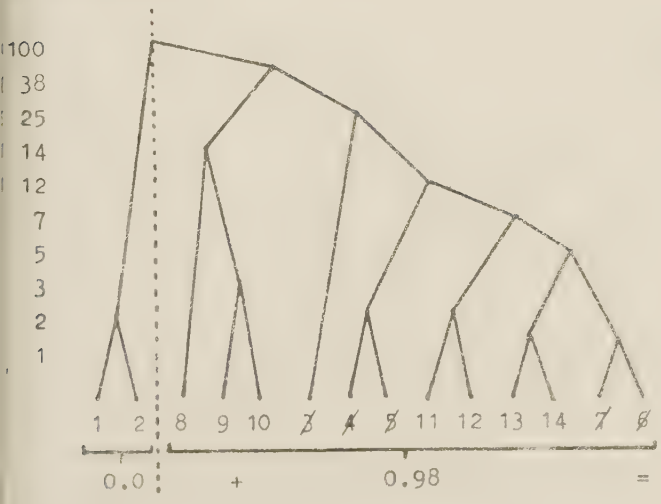
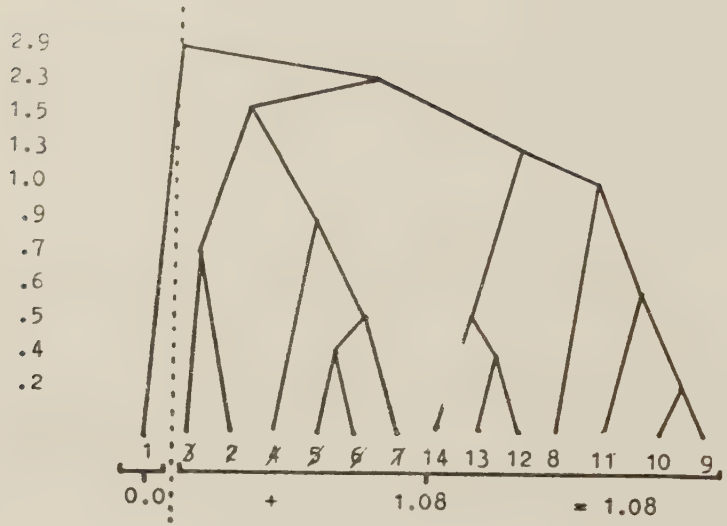



Abb. : Zur Erläuterung [illegible] Seite 84.

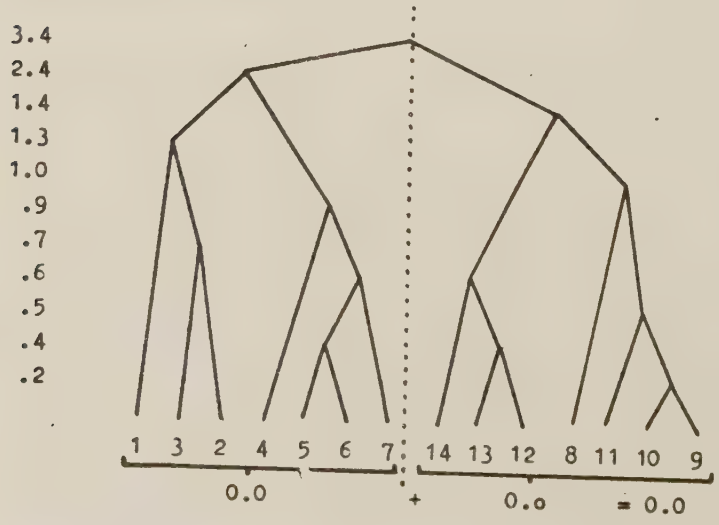
ROSSBACH

Maximum-Methode




g.-o.-f.: 2

Schwerpunkt-Methode

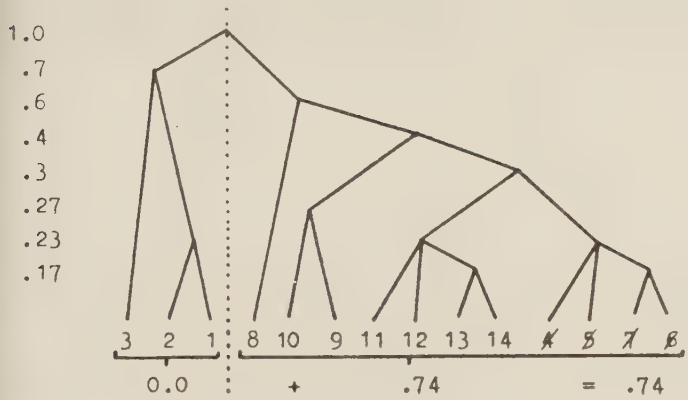



g.-o.-f.: 0.0

Abb. : Zur Erläuterung siehe Kommentar auf Seite 84.

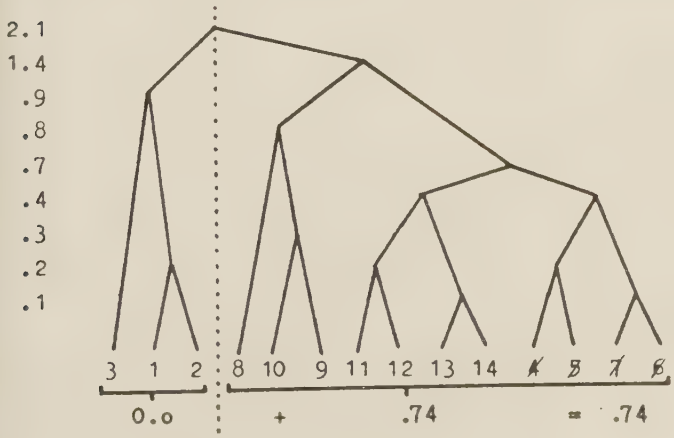
INFOTWO

Minimum-Methode




g.-o.-f.: 664

Gruppendurchschnitt-Methode

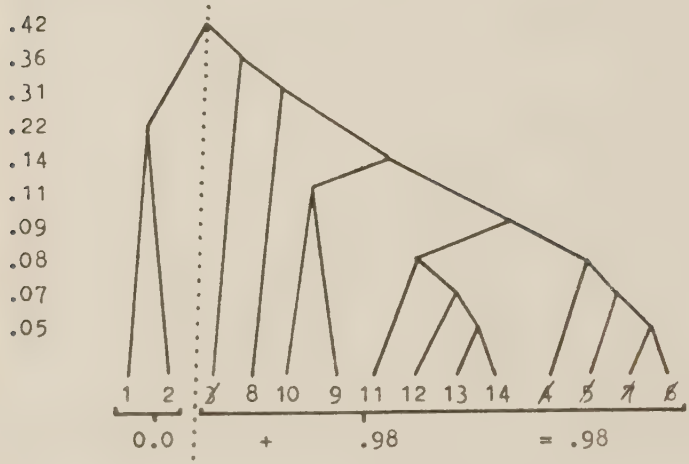



g.-o.-f.: 592

Abb. : Zur Erläuterung siehe Kommentar auf Seite 84.

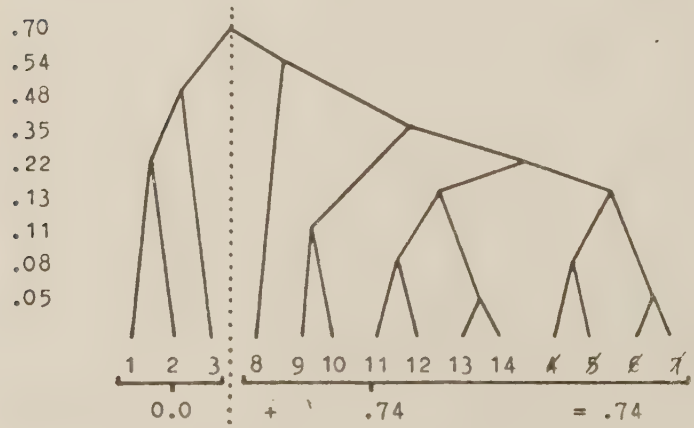
INFOTHRY

Minimum-Methode




g.-o.-f.: 774

Gruppendurchschnitt-Methode




g.-o.-f.: 672

Abb. : Zur Erläuterung siehe Kommentar auf Seite 84.

### 6.1.3 Die Bewertung der Baumgraphen des ersten Datensatzes

Glücklicherweise hat es sich zufällig so ergeben, dass die Bäume dieses Datensatzes gut dazu geeignet sind, zu verdeutlichen, welche ihrer strukturellen Eigentümlichkeiten, die zu ihrer Bewertung bestimmten Kriterien (erster und zweiter Test nach Abschnitt 5.3 und 5.5) erfassen. Das ist vor allem darauf zurückzuführen, dass einzelne Bäume darunter sind, die durch die beiden Tests sehr unterschiedlich bewertet werden. Denn selbstverständlich zeigt sich an solchen Beispielen besonders klar, worauf sich das jeweilige Kriterium stützt.

Unter den abgebildeten Bäumen ist nur einer, der von beiden Tests als "absolut" oder "genau richtig" eingestuft wird: Die Kombination ROSSBACH/Schwerpunktsmethode führt als einzige auf einen Baum mit einer Diskrepanz von Null zur theoretischen Aufteilung. Gleichzeitig versammelt dieser Baum unter seinen zwei Hauptästen Aufteilungen, die eine strukturell vollständig homogene Gruppe bilden. Verallgemeinernd lässt sich sagen, dass in einem Fall, den man auch intuitiv als die richtige Lösung einstuft, die Tests übereinstimmend ebenfalls zu diesem Urteil kommen.

Das ist bei der Bewertung von "Graden der Falschheit" durchweg anders. Der in dieser Hinsicht krasseste und damit lehrreichste Fall ist der Baum ROSSBACH/Maximum-Methode. Er hat die zweitbeste Bewertung hinsichtlich seines Goodness-of-fit mit der theoretischen Zweiteilung (252) und ist am schlechtesten bezüglich Homogenität (1.08). Letzteres ist nicht erstaunlich; sind doch unter dem Ast sechs Aufteilungen der einen Gruppe mit sämtlichen der anderen vereinigt. Interessanter ist jedoch die Frage, wieso es trotz der sechs Fehlklassifikationen zu einem so guten Abschneiden im ersten Test kommt. Nun, das ist durch einen Blick auf die Struktur des Baumes sofort beantwortet. Die sechs zusammengehörigen Aufteilungen sind getrennt von den übrigen unter demselben Teilast vereinigt. Schon ein um eine Stufe niedrigerer Schnitt durch den Baum ergibt drei homogene Gruppen, darunter eine einelementige, bestehend aus dem Ausreisser $P_1$. So gesehen handelt es sich eigentlich

um einen sehr guten Baum. Man könnte daraus den Schluss ziehen, der erste Test sei dem zweiten generell vorzuziehen. Das ist aber verfehlt, denn eine Inspektion feinerer Aufteilungen führt, wie das Beispiel lehrt, auch bei Anwendung des zweiten Tests zur Aufklärung solch irreführender Bewertungen.

Eine weitere Möglichkeit, sich über die Auswirkung struktureller Unterschiede auf die Bewertung, Klarheit zu verschaffen, besteht im Vergleich mehrerer Bäume, die von dem einen Mass gleich, von dem anderen verschieden beurteilt werden. Die Bäume INFOTWO/Gruppendurchschnitt und INFOTHRY/Gruppendurchschnitt sind dafür ein Beispiel. Die Inhomogenität beider beträgt .74, aber der Baum der INFOTHRY-Distanzen schneidet mit 672 gegenüber 592 im goodness-of-fit Test erheblich schlechter ab. Das ist nicht gleich verständlich, denn die beiden Bäume sind strukturell fast identisch. Eigentlich gibt es nur zwei Unterschiede: $P_1$ und $P_2$, obwohl Elemente derselben theoretischen Teilmenge, haben sie beim schlechter eingestuften Baum eine grössere Distanz zueinander und gleichzeitig hat $P_8$ zu keinem Element der eigenen Gruppe eine niedrigere Distanz, als zu allen anderen ihres Hauptastes. Es leuchtet ein, dass das HUBERTsche Distanzmass, das ja dem Goodness-of-fit-Test zugrundeliegt, Inkonsistenzen in genau dieser Weise bewertet: Ein Baum ist danach umso besser, je mehr Distanzen zwischen Aufteilungen aus d e r s e l b e n theoretischen Teilmenge k l e i n e r sind als Distanzen zwischen Elementen aus verschiedenen.

Man könnte auf die Idee kommen, gleichsam als dritten Test schlicht nach der Anzahl der falsch klassifizierten Aufteilungen unter den beiden Hauptästen zu fragen. Bei den Bäumen dieses Datensatzes kommen drei Fälle vor: sechs, fünf, vier und null Fehlklassifizierungen. Auch wenn man letzlich an nichts anderem interessiert ist, als möglichst viele richtige Klassifikationen zu erhalten, mechanisch angewendet ist dieser intuitiv so einleuchtende Test zu oberflächlich. Ein einzelner Aussenseiter unter den klassfizierten Aufteilungen kann, wie das Beispiel des Baumgraphen ROSSBACH/Maximum-Methode zeigt, zur Unterbewertung eines Baumes führen, der sonst eine klare Separation der erwarteten Gruppen liefert. Die relativ enge Zusammenfassung theoretisch zusammengehöriger Elemente bleibt unberücksichtigt, nur weil sie innerhalb eines

grösseren Clusters erfolgte, das auch noch fremde Elemente enthält.

Im übrigen ist es wichtig, sich vor Augen zu halten, dass dieses Kriterium (der Anzahl der Fehlklassifikationen) ebenso wie der erste Test, der die Diskrepanz des Baumes zur theoretischen Zweiteilung misst, die Kenntnis der "richtigen" Aufteilung der Menge der Aufteilungen voraussetzt. Unter realen Anwendungsbedingungen ist also nur über den zweiten Test (Inhomogenität der gebildeten Teilgruppe) Aufschluss über die Qualität des Ergebnisses zu erlangen. Da man im Ernstfall natürlich verschieden feine Schnitte durch den erhaltenen Baumgraphen, also verschieden feine Aufteilungen der Menge der Aufteilungen, miteinander vergleichen wird, sind Fehleinschätzungen durch diesen Test für Inhomogenität praktisch ausgeschlossen - sogar auch dann, wenn nicht die Möglichkeit besteht, durch Inspektion des Baumes eine entsprechende Vermutung zu begründen.(So hat beispielsweise im Fall des Baumes ROSSBACH/ Maximum-Methode schon die nächst feinere Aufteilung in

$$\{(1),(2,3,4,5,6,7),(8,9,10,11,12,13,14)\}$$

eine Inhomogenität von Null). Ein höherer Wert für Inhomogenität bei der Teilung durch den obersten Gabelungspunkt bedeutet also nicht notwendig, dass der Baumgraph schlecht ist (das kann erst aufgrund der Kenntnis weiterer, feinerer Aufteilungen entschieden werden), wohl aber ist ein guter Wert hinreichend für ein positives Urteil. Da sich etwas entsprechendes vom ersten Test nicht sagen lässt, wird er hier nur eingesetzt um in Fällen, wo die oberste, d.h. gröbste Aufteilung nach dem zweiten Test ein verkehrtes Bild geben würde, als Korrektiv zu wirken. Das ist - wenn man die grundsätzliche Vorrangigkeit des Inhomogenitätstestes im Auge behält - fast immer unproblematisch, weil die theoretische Zweiteilung ja ihrerseits auf Homogenität hin konzipiert ist.

Der vollständige Vergleich der zwanzig Ergebnisse anhand der Tabelle 1 S.93 bringt nichts wesentlich Neues. Wir sehen, dass drei Verfahrenskombinationen, die sämtlich auf dem Distanzmass ROSSBACH gründen, einheitlich von beiden Tests als "richtig" eingestuft werden. Die beiden anderen Clustermethoden liefern mit demselben

Mass ein Ergebnis, das immer noch besser als die Resultate aller anderen Kombinationen ist. Die Ergebnisse mit dem Mass PAIRBONDS sind für alle Clusterverfahren die schlechtesten. Diese Tendenz, genau wie die über alle Distanzmasse hinweg vergleichsweise geringe Leistung der Minimum-Methode, zeigt sich auch bei den folgenden Datensätzen. Diese Gleichförmigkeit stützt erheblich unsere These über die Bedeutung dieses ersten Datensatzes.

Verfahrenskombinationen, die an dieser "reinen" Plasmode mit ihren denkbar günstigen Voraussetzungen zur Auffindung der vorgegebenen Gruppierung bereits scheitern, dürften für die Analyse weniger klar strukturierter Daten erst recht ungeeignet sein.

Je nachdem, wie eng man nun den Begriff des "Scheiterns" bzw. der "Bewährung" fasst, wird man das Ergebnis der Analyse des ersten Datensatzes etwas anders zusammenfassen:

Unter Zugrundelegung eines weiten Begriffs von Bewährung, wird man lediglich von den Distanzmassen PAIRBONDS und von den Klassifikationsverfahren die Minimum-Methode als gescheitert ansehen. Legt man dagegen einen strengeren Massstab an, so muss man zu dem Ergebnis kommen, dass nur Baumgraphen auf der Basis des Distanzmasses ROSSBACH die vorgegebene Zweiteilung hinreichend genau reproduzieren. Clusteranalysen nach Gruppendurchschnitt-, Schwerpunkt- und WARD-Methode tun dies exakt, die beiden anderen mit geringfügigen Abweichungen. Alle anderen Methodenkombinationen versagen vor der ihnen gestellten Aufgabe.

Tabelle : 1

Die Bewertung der Baumgraphen des ersten Datensatzes

| | INHOMOGENITÄT | GOODNESS-OF-FIT |
|---|---|---|
| PAIRBONDS | | |
| Minimum-Methode | .98 | 75o |
| Maximum-Methode | .98 | 734 |
| Gruppendurchschnitt-Methode | .98 | 8o2 |
| Schwerpunkt-Methode | .98 | 734 |
| Ward-Methode | .98 | 734 |
| ROSSBACH | | |
| Minimum-Methode | .74 | 336 |
| Maximum-Methode | 1.o8 | 252 |
| Gruppendurchschnitt-Methode | o.o | o |
| Schwerpunkt-Methode | o.o | o |
| Ward-Methode | o.o | o |
| INFOTWO | | |
| Minimum-Methode | .74 | 664 |
| Maximum-Methode | .74 | 528 |
| Gruppendurchschnitt-Methode | .74 | 572 |
| Schwerpunkt-Methode | .74 | 528 |
| Ward-Methode | .74 | 528 |
| INFOTHRY | | |
| Minimum-Methode | .98 | 774 |
| Maximum-Methode | .74 | 592 |
| Gruppendurchschnitt-Methode | .74 | 672 |
| Schwerpunkt-Methode | .74 | 592 |
| Ward-Methode | .74 | 56o |

## 6.2 Der zweite Datensatz

### 6.2.1 Die Entstehung des zweiten Datensatzes

Der zweite Datensatz ist in Anlehnung an den ersten konstruiert worden. Er besteht aus ebensovielen Aufteilungen derselben Anzahl von Elementen. Jede Aufteilung des ersten Datensatzes ist geringfügig verändert worden, sodass schliesslich keine mehr mit dem ursprünglich zu ihrer Konstruktion verwendeten Baumgraphen vollständig übereinstimmt. Alle weisen aber noch eine grössere Ähnlichkeit zu ihm auf. Auf diese Art soll die unter natürlichen Erhebungsbedingungen normalerweise stets vorhandene Fehlerstreuung simuliert werden.

Aufteilungen des zweiten Datensatzes (Tabelle 2) siehe Seite 95.

Tabelle 2 : Die Aufteilungen des zweiten Datensatzes

Q1 = {(3,6),(1,2,7,8,12),(4,5,9,1o,11)}

Q2 = {(3),(1,2),(4,6,12),(5,9,1o,11),(7,8)}

Q3 = {(3),(6),(1),(2,4),(5,9,11,12),(7,8,1o)}

Q4 = {(3,7),(6,8),(1,1o),(2),(4,12),(5),(9),(11)}

Q5 = {(3),(6),(1,2),(12),(4),(5),(9),(11),(7,8,1o)}

Q6 = {(3),(6),(1),(2,12),(4),(5),(9),(11),(7,8),(1o)}

Q7 = {(3,6,1o),(1,2,4),(5,12),(8,9,1o,11)}

Q8 = {(1,4,5,11),(6,7),(2,3,9,8,1o,12)}

Q9 = {(1),(3,4,11),(2,6,7,5,9),(8,1o,12)}

Q1o= {(1),(4,11),(2,3,8,9),(5,6,7,1o,12)}

Q11= {(1),(4),(11),(3),(2,5,6,7),(8),(9,1o,12)}

Q12= {(1),(4,9),(11),(3),(2,7),(5,6),(8,1o),(12)}

Q13= {(1),(4),(11),(3),(2,6),(7),(5),(9),(8),(1o,12)}

Q14= {(1,5,11),(2,3,4,6,8,9,1o,12),(7)}

### 6.2.2 Die Analyse des zweiten Datensatzes

Tabelle 3 zeigt die Ergebnisse des zweiten Datensatzes, d.h. die "empirischen" Zweiteilungen und die quantitativen Bewertungen der zwanzig Baumgraphen hinsichtlich ihrer Diskrepanz zur theoretischen Aufteilung und der Inhomogenität ihrer beiden Hauptäste. Der Gesamteindruck entspricht dem des ersten Datensatzes.: Alle auf PAIRBONDS gegründeten Baumgraphen werden ziemlich schlecht beurteilt - und zwar nach beiden Kriterien. Wieder gehören alle Bäume, die sich vom Distanzmass ROSSBACH herleiten zu den Besten. Etwas günstiger schneiden diesmal alle mit der Maximum-Methode gewonnene Baumgraphen ab. Das ist durchaus im Sinne der Erwartungen, die man dieser Methode allgemein entgegenbringt. Es fällt auf, dass diesmal derjenige Baum, der mit den Clustern seiner beiden Hauptäste genau die theoretisch erwartete Aufteilung der Menge erzeugt (ROSSBACH/Schwerpunkt-Methode) und daher den Diskrepanzwert Null erhält, hinsichtlich der Homogenität nicht ebenfalls den ersten Platz einnimmt. Stattdessen erhalten drei andere Bäume (ROSSBACH/Max.-M., INFOTWO/Max.-M., INFOTHRY/Max.-M.) eine mit 1.29 gegenüber 1.32, wenn auch nur geringfügig bessere Bewertung. Dabei sind die Ergebnisse der beiden zuletzt genannten Methodenkombinationen im Goodness-of-fit sogar erheblich schlechter (177 bzw. 213) als ein weiterer Graph (ROSSBACH/WARD-M.), der nur eine Diskrepanz von 78 zur theoretischen Gruppierung aufweist.

Es liegt nahe, dieses zunächst überraschende Ergebnis durch die anschauliche Gegenüberstellung der Baum-Graphen zu erhellen. Um der besseren Vergleichbarkeit willen, haben wir von den drei Graphen mit minimaler Inhomogenität denjenigen ausgewählt, der ebenfalls auf der Basis des Distanzmasses ROSSBACH gebildet wurde. Damit liegen uns drei in der Struktur sehr ähnliche Bäume in der Abb. 12 vor.

Die Zweiteilung des ersten stimmt mit der theoretsichen genau überein. Bezüglich der Homogenität des rechten Teilastes wird er jedoch deutlich (mit .47 statt .59) vom folgenden Baum übertroffen. Dieser Gewinn an Homogenität des einen Astes, durch den Transfer der Aufteilung Q12 bewirkt,mußte mit einer

Zunahme der Inhomogenität des anderen Clusters von .73 auf .82 erkauft werden. Trotzdem verbleibt ein geringer Nettogewinn an Homogenität, der Unterschied zwischen 1.29 und 1.32 im Gesamtwert.

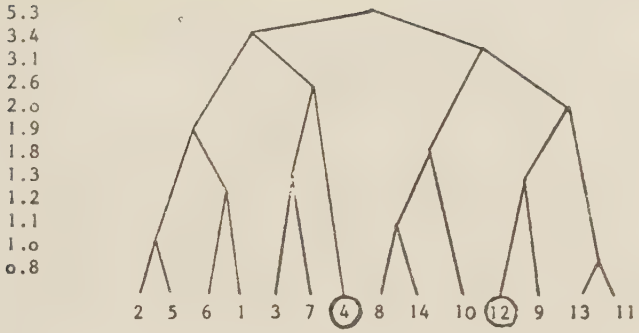
Der obere Graph weist ein zweites kritisches Element auf: Die Aufteilung Q4 ist von allen am lockersten, nämlich über den für ein einzelnes Element höchsten Knoten mit ihrem Cluster verbunden. Tatsächlich gibt es eine Aufteilung (ROSSBACH/WARD, unterer Graph) die Q4 dem anderen Cluster zuordnet. Nur dass in diesem Fall der beträchtliche Gewinn an Homogenität bei dem Cluster, aus dem Q4 ausscheidet (von .73 auf .40) durch den Zuwachs beim anderen Cluster weit übertroffen wird (von .59 auf 1.03). Q4 ist also i n s g e s a m t eine schlecht integrierbare Aufteilung. Das kommt darin zum Ausdruck, dass Q4 als einziges Element bei zwei Bäumen (ROSSBACH/Min.-M. und ROSSBACH/Gruppend.-M.) sogar von allen übrigen Elementen abgetrennt wird. Der Wert für die Inhomogenität der gesamten Restgruppe (alle Elemente ausser Q4) ist mit 1.67 erstaunlich niedrig, wenn man bedenkt, dass die Ausgliederung der vier Elemente Q1, Q2, Q3, Q7 eine Restgruppe mit der Inhomogenität 1.91 ergibt (INFOTWO/Schwerp.-M.). Die mittlere und homogenste Aufteilung der Abbildung fasst Q4 und Q12 unter dem höchsten Knoten zusammen, der zwei einzelne Elemente miteinander verbindet und verküpft sie unmittelbar unter der obersten Gabelung mit ihrem Cluster. Das zeigt ebenfalls ihre Aussenseiterposition.

Der Vergleich der drei Bäume von Abbildung 12 zeigt, welche Informationen derartigen Graphen bei Berücksichtigung der Inhomogenität ihrer Hauptäste zu entnehmen ist. Vor allem erweist sich, dass das Entfernen von nur locker mit ihrem Cluster verbundenen Aufteilungen einen erheblichen Homogenitätszuwachs zur Folge hat. Die schwache Verknüpfung eines Elementes mit dem einen Cluster ist aber natürlich auf keinen Fall gleichbedeutend damit, dass seine Eingliederung in ein anderes Cluster zu einer günstigeren Gesamtbewertung führt.
Andeutungsweise zeigt sich an diesem künstlichen Datensatz mit schwach ausgeprägter "noise"-Komponenten auch schon, was beim empirischen noch deutlicher wird: die Übereinstimmung eines Baumes mit der theoretischen Unterteilung ist nicht gleichbedeutend mit maximaler Homogenität der betreffenden Gruppierung.

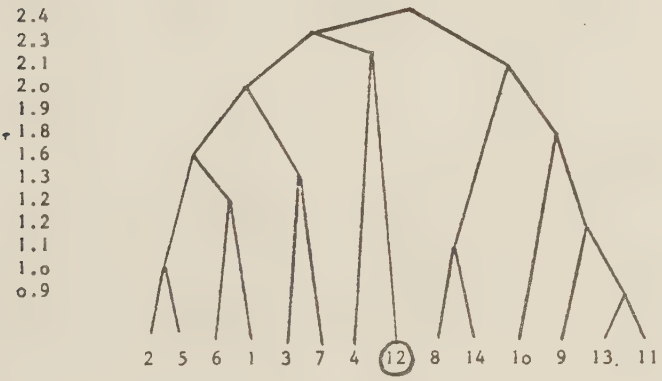
| | | Inhomogenität | | Goodness-of-fit |
|---|---|---|---|---|
| PAIRBONDS | | | | |
| Min, Max, GrD, SPu, Wrd | (1,2,3,4,5,6,7,9,1o,11,12,13) (8,14) | 1.91 .o3 | 1.94 | 863 647 876 639 624 |
| ROSSBACH | | | | |
| Min | (1,2,3,5,6,7,8,9,1o,11,12,13,14) (4) | 1.67 o.o | 1.67 | 252 |
| Max | (1,2,3,4,5,6,7,12) (8,9,1o,11,13,14) | .82 .47 | 1.29 | 48 |
| GrD | (1,2,3,5,6,7,8,9,1o,11,12,13,14) (4) | 1.67 o.o | 1.67 | 252 |
| SPu | (1,2,3,4,5,6,7) (8,9,1o,11,12,13,14) | .73 .59 | 1.32 | o |
| Wrd | (1,2,3,5,6,7) (4,8,9,1o,11,12,13,14) | .4o 1.o3 | 1.43 | 78 |
| INFOTWO | | | | |
| Min | (1,2,3,4,5,6,7,9,11,12,13) (8,1o,14) | 1.5o .18 | 1.68 | 797 |
| Max | (1,2,3,4,5,6,7,12) (8,9,1o,11,13,14) | .82 .47 | 1.29 | 177 |
| GrD | (1,2,3,4,5,6,7,9,11,12,13) (8,1o,14) | 1.5o .18 | 1.68 | 699 |
| SPu | (1,2,3,7) (4,5,6,8,9,1o,11,12,13,14) | .24 1.91 | 2.15 | 473 |
| Wrd | (1,2,3,4,5,6,7,9,11,12,13) (8,1o,14) | 1.5o .18 | 1.68 | 561 |
| INFOTHRY | | | | |
| Min | (1,2,3,4,5,6,7,9,1o,11,12,13) (8,14) | 1.91 .o3 | 1.94 | 9o3 |
| Max | (1,2,3,4,5,6,7,12) (8,9,1o,11,13,14) | .47 .81 | 1.29 | 213 |
| GrD | (1,2,3,4,5,6,7,9,1o,11,12,13) (8,14) | 1.91 .o3 | 1.94 | 87o |
| SPu | (1,2,3,8,14) (4,5,6,7,9,1o,11,12,13) | 1.o6 1.17 | 2.24 | 633 |
| Wrd | (1,2,3,4,5,6,7,9,1o,11,12,13) (8,14) | 1.91 .o3 | 1.94 | 666 |

Tabelle 3 : Ergebnisse des zweiten Datensatzes, Aufteilung der Menge der Aufteilun anhand der obersten Gabelung des jeweiligen Baumgraphen, Inhomogenität der so gebi ten Gruppen und davon die Summe, Goodness-of-fit des Baums mit der theoretischen Z teilung

ROSSBACH/Schwerpunkt-M.




ROSSBACH/Maximum-M.




ROSSBACH/Ward-M.

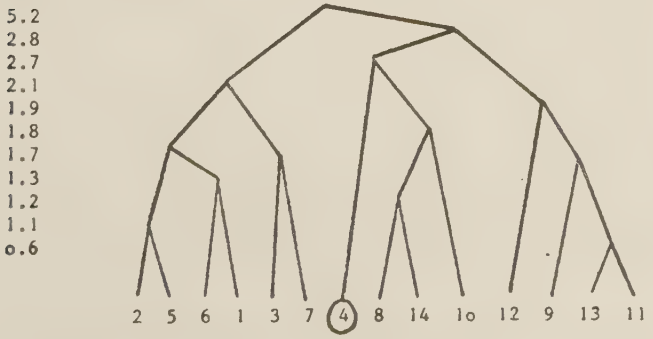



Abb. 12: Drei Baumgraphen des zweiten Datensatzes. Siehe die Erläuterungen S.

## 6.3 Der dritte Datensatz

### 6.3.1 Die Entstehung des dritten Datensatzes

Mit der dritten Untersuchung wurde die differentielle Analyse von Sorting-Daten nun erstmalig unter normalen Anwendungsbedingungen durchgeführt. Um der Aussagekraft des Experimentes willen, wurde bewusst davon abgesehen, besonders günstige Daten im Sinne der Methode vorzugeben. (Dies wäre schon deshalb wenig sinnvoll, weil damit den vorangehenden Experimenten keine neue Information hinzugefügt würde). Gleichzeitig sollten aber auch die Voraussetzungen eines Methodenexperimentes erfüllt sein. Das hiess vor allem, dass eine Validierungsgrundlage, anhand derer der Erfolg des in Erprobung befindlichen Instrumentariums abgeschätzt werden kann, vorhanden sein musste. In Bezug auf das hier anstehende Problem bedeutete das, dass im Vorwege sichergestellt wurde, welche reale Gruppierung methodisch rekonstruiert werden sollte.

Die Unterscheidbarkeit der Aufteilungen nach zwei im Voraus bekannten Klassen wurde dadurch herzustellen versucht, dass in ihrer Einstellung gegensätzliche (wie gemeinhin angenommen wird) Extremgruppen von Versuchspersonen herangezogen wurden: "linke" und "rechte" Studenten. Die Operationalisierung bestand in der Auswahl von je 16 Personen, die entsprechend orientierten studentischen Vereinigungen angehörten: bei "Rechten" war das der "Ring Christlich-Demokratischer Studenten" oder eine Burschenschaft, bei "Linken" die "Kommunistische Studenten-Gruppe" und der "Marxistische Studentenbund Spartakus".

Beiden Gruppen von Versuchspersonen wurden dieselben Items, nämlich zwanzig für die politische Auseinandersetzung kritische Begriffe zum Sortieren vorgelegt. Erfahrungsgemäss sind derartige Begriffe für poltische Gegner mindestens konnotativ, also hinsichtlich ihrer affektiven, assoziativen Beiklänge, möglicherweise aber sogar denotativ verschieden.

Die wichtigste Voraussetzung dieser Erhebung bestand in der

Annahme, dass diese Auffassungsunterschiede spezifisch für die so bestimmten Extremgruppen sind und in der Art und Weise, wie die Versuchspersonen die Begriffe sortieren, deutlich genug zum Ausdruck kommen würden. Es machte den Realismus dieses Methodenexperimentes aus, dass über den Grad der Separation und Homogenität der Aufteilungen der theoretischen, d.h. der durch die Operationalisierung bestimmten Extremgruppen, vor der Durchführung der Untersuchung keinerlei Wissen vorhanden war. Im Idealfall, nämlich bei vollständiger Homogenität jeder der beiden Teilgruppen würde der auf der Grundlage ihrer jeweiligen Aufteilungen konstruierte Baum der Begriffe die, von jeder Person einer Gruppe geteilte, "rechte" beziehungsweise "linke" semantische Struktur widerspiegeln.

Die Begriffe waren die folgenden:

| | | | |
|---|---|---|---|
| Widerstand | Minorität | Ordnung | Macht |
| Entwicklungshilfe | Unterdrückung | Ideologie | Eigentum |
| Monopol | Presse | Manipulation | Agitation |
| Demokratie | Streik | Klassenkampf | Arbeiterbewegung |
| Freiheit | Volksinteresse | Gerechtigkeit | Fortschritt |

Die Wörter waren, jedes für sich, auf kleine Kärtchen geschrieben und wurden den Versuchspersonen in zufälliger Reihenfolge (als gut gemischter Kartenstoss) mit dieser Anweisung überreicht:

> "Ordnen Sie die Begriffe auf den Kärtchen und bilden Sie Gruppen, indem Sie jeweils diejenigen Begriffe zusammenlegen, die nach Ihrem Empfinden ähnlich sind. Sie können so viele und so grosse Gruppen bilden, wie Sie wollen. Die Anordnung innerhalb einer Gruppe bleibt unberücksichtigt."

### 6.3.2 Die Analyse des dritten Datensatzes

Wie die Tabelle 4 zeigt, fällt die Bewertung dieser zwanzig Baumgraphen durch die beiden Kriterien noch stärker auseinander als beim vorigen Datensatz. Das Ergebnis von INFOTHRY/Maximum-Methode zeigt die beste Übereinstimmung mit der vorgängigen Einteilung der Versuchspersonen in "rechte" und "linke", wird aber hinsichtlich Homogenität von vier anderen Baumgraphen übertroffen. An der Spitze steht bezüglich dieses Kriteriums der Graph der Methodenkombination ROSSBACH/WARD.

Im Vergleich zu den beiden anderen Datensätzen verteilt sich der relative Erfolg der verschiedenen Kombinationen diesmal etwas anders. Die auf der Distanz ROSSBACH basierenden Bäume gehören nicht wie sonst sämtlich zu den Ersten. (Immerhin ist das beste Ergebnis überhaupt, wie gesagt, wieder unter ihnen zu finden). Einigermassen irritierend ist aber die Tatsache, dass die nach ROSSBACH/WARD nächstbesten Aufteilungen

PAIRBONDS/Schwerp.-M. {(2,3,4,5,6,7,9,11,12,13,15,16,20,23,24,25,29),
(1,8,10,14,17,18,19,21,22,26,27,28,30,31,32,)}

und

INFOTWO/Schwerp.-M. {(1,3,4,5,6,8,10,12,13,14,15,18,19,20,26),
(2,7,9,11,16,17,21,22,23,24,25,27,28,29,30,31,32)}

von der wenig besseren, optimalen Aufteilung (ROSSBACH-WARD) - aber auch von einander - so stark verschieden sind. Dieser Umstand ist jedoch kein Grund, am Wert der Methode der inversen Clusteranalyse überhaupt zu zweifeln. Da im Fall der nachweislich realisierten Zweiteilung (wie bei den ersten Datensätzen) übereinstimmend die Möglichkeit ihrer erfolgreichen Rekonstruktion demonstriert wurde - bei Anlegung strenger Kriterien mindestens für die Methodenkombination ROSSBACH/WARD - liegt es nahe, dies Ergebnis so zu interpretieren:

Erwartungsgemäss wurde die relativ beste Gruppierung wieder durch Anwendung von ROSSBACH/WARD erzielt. Die erheblichen Un-

terschiede der Zusammensetzung der drei besten Aufteilungen sind nicht allein aus der Natur des Agglomerationsprozesses zu erklären. (Cluster höherer Stufe und damit vor allem die Zweiteilung auf der Grundlage des obersten Gabelungspunktes, sind das Ergebnis vorangehender Vereinigungen von Clustern niedrigerer Stufen. Wegen der Irreversibilität dieses Prozesses und weil die Klassifikationsmethoden - jedenfalls, wenn die Daten nicht ultrametrisch sind - entsprechend ihrem Kriterium jeweils andere Cluster bilden, akkumulieren sich mit jeder Stufe derartige Unterschiede). Auch auf der Ebene der Distanzmasse muss schon ein erhebliches Mass an Verschiedenheit vorgelegen haben, denn sonst wären ja die Zweiteilungen nach derselben Klassifikationsmethode (Schwerpunkt-M.) ähnlicher ausgefallen. Wir müssen demnach davon ausgehen, dass die mit der Auswahl der Versuchspersonen und Items angestrebte Deutlichkeit der Dichotomie faktisch nicht hinreichend realisiert gewesen ist.

Dieser Versuch zur Exhaustion des im Wesentlichen wegen der nur mässigen Übereinstimmung mit den beiden ersten nicht voll befriedigenden dritten empirischen Tests basiert auf der Annahme, dass es in diesem Fall bereits von der "wahren Struktur" der Daten hier d i e richtige Zweiteilung gar nicht gibt. Der Versuch trotzdem eine derartige Dichotomie dem Datenkörper zu oktroyieren, führt dann zu absolut gesehen nicht sehr homogenen Zweiteilungen von ziemlich unterschiedlicher Zusammensetzung. (Nach dem Beispiel der drei Punkte, die durch eine Trennlinie auf der von ihnen bestimmten Ebene in beliebiger Weise aufgeteilt werden können - es sei denn sie lägen auf einer Geraden). Die weitere Absicherung dieser Erklärung hätte folgerichtig in dem Nachweis zu bestehen, dass die Inhomogenität der besten Zweiteilungen dieses dritten Datensatzes im Vergleich zu den besten Lösungen anderer von der Struktur geeigneterer Datensätze (wie dem ersten und zweiten) deutlich grösser ist. (Tabelle 4 - s. Seite 104)

Tabelle 4:

Die Bewertung der Baumgraphen des dritten Datensatzes

| | Inhomogenität | Goodness-of-fit |
|---|---|---|
| PAIRBONDS | | |
| Minimum-Methode | - | 27492 |
| Maximum-Methode | 5.46 | 2o931 |
| Gruppendurchschnitt-Methode | 6.oo | 25736 |
| Schwerpunkt-Methode | 4.49 | 14472 |
| Ward-Methode | 5.16 | 24985 |
| ROSSBACH | | |
| Minimum-Methode | - | 28526 |
| Maximum-Methode | 5.75 | 191o7 |
| Gruppendurchschnitt-Methode | - | 17165 |
| Schwerpunkt-Methode | 5.o6 | 14637 |
| Ward-Methode | 4.27 | 8616 |
| INFOTWO | | |
| Minimum-Methode | - | 283o1 |
| Maximum-Methode | 4.59 | 8452 |
| Gruppendurchschnitt-Methode | 5.23 | 16951 |
| Schwerpunkt-Methode | 4.47 | 14182 |
| Ward-Methode | 4.51 | 1o84o |
| INFOTHRY | | |
| Minimum-Methode | - | 29o26 |
| Maximum-Methode | 4.69 | 4471 |
| Gruppendurchschnitt-Methode | - | 25o87 |
| Schwerpunkt-Methode | 4.93 | 2o1o5 |
| Ward-Methode | 4.6o | 11o14 |

Denkbar wäre es, darüber durch eine geignete Normierung des Inhomogenitätsmasses Kenntnis zu erhalten. Da sich der Wert für die Inhomogenität eines Clusters aus der Summe der Diskrepanzen zwischen den Aufteilungen dieses Clusters und dem auf ihrer Grundlage konstruierten Item-Baum ergibt, bietet es sich an, diesen Wert einfach durch die Anzahl der Summanden zu teilen. Dieser so relativierte (bzw. Durchschnitts-) Inhomogenitätswert wurde für alle zwanzig Ergebnisse der drei Datensätze berechnet. Allein die Tatsache, dass die beste Lösung des dritten Datensatzes danach mit .133 gegenüber .092 deutlich schlechter als das Optimum des zweiten eingestuft wird, lässt vermuten, dass diese Art der Normierung für Vergleiche zwischen Datensätzen mit verschiedenen Parametern sinnvoll sein dürfte.

Trotz dieser Schwäche bei der methodischen Ermittlung der besten Zweiteilung der Versuchspersonen in "Linke" und "Rechte", wurde auf der Basis der homogensten Dichotomie die Konstruktion des semantischen Gefüges der politischen Begriffe für beide Teilgruppen durchgeführt. (Abb. 14 und Abb. 15).

Wegen der Undeutlichkeit dieser Gruppierung (ROSSBACH/WARD), Abb. 13) verbietet sich eine allzu tiefschürfende Interpretation der beiden Baumgraphen. Aber die zum Teil frappierende Plausibilität der Begriffsstrukturen fordert doch zu ein paar Anmerkungen heraus:
Als insgesamt am engsten zusammengehörig werden von den "Rechten" "Ordnung" und "Freiheit" empfunden ! Die beiden Begriffe finden sich bei ihnen in einem grösseren Cluster zusammen mit "Gerechtigkeit" und "Demokratie". Die"Linken" dagegen verbinden "Ordnung" mit "Macht", die dem eng assoziierten Paar "Entwicklungshilfe" und "Unterdrückung" (!) beigesellt sind. Das steht offenbar in scharfem Gegensatz zur Auffassung der "Rechten" über "Entwicklungshilfe", die sie mit "Fortschritt" verbinden. In dieser Weise lässt sich fortfahren: "Linke" bilden mit "Eigentum" ein sicherlich negativ getöntes Cluster aus "Monopol" und "Ideologie", während die "Rechten" "Eigentum"

in die oben erwähnte durchgehend positiv bewertete Klasse einordnen, und so weiter.

Die unzureichend klare Strukturiertheit des dritten Datensatzes legt, wie gesagt, eine gewisse Zurückhaltung gegenüber weitergehenden Deutungen nahe. Prinzipiell lässt sich jedoch eine Auswertung denken, die intensiveren Gebrauch von der vorliegenden Information macht. So könnten beispielsweise Gruppen von Aufteilungen beziehungsweise Versuchspersonen verschiedener Homogenität gebildet und ihre zugehörigen Item- also beispielsweise Begriffsstrukturen erzeugt und miteinander verglichen werden. Als Methoden zur Darstellung der zwischen den Items bestehenden Beziehungen kommen nicht nur die in dieser Arbeit verwendeten Hierarchischen Clusteranalysen in Frage, sondern ebenso Nicht-hierarchische Klassifikationsverfahren und verschiedene Techniken der Multidimensionalen Skalierung. Die Vergleiche zwischen den erhaltenen Itemstrukturen können quantitativ und qualitativ sein. Das Feld der Möglichkeiten ist damit nur in Umrissen beschrieben.

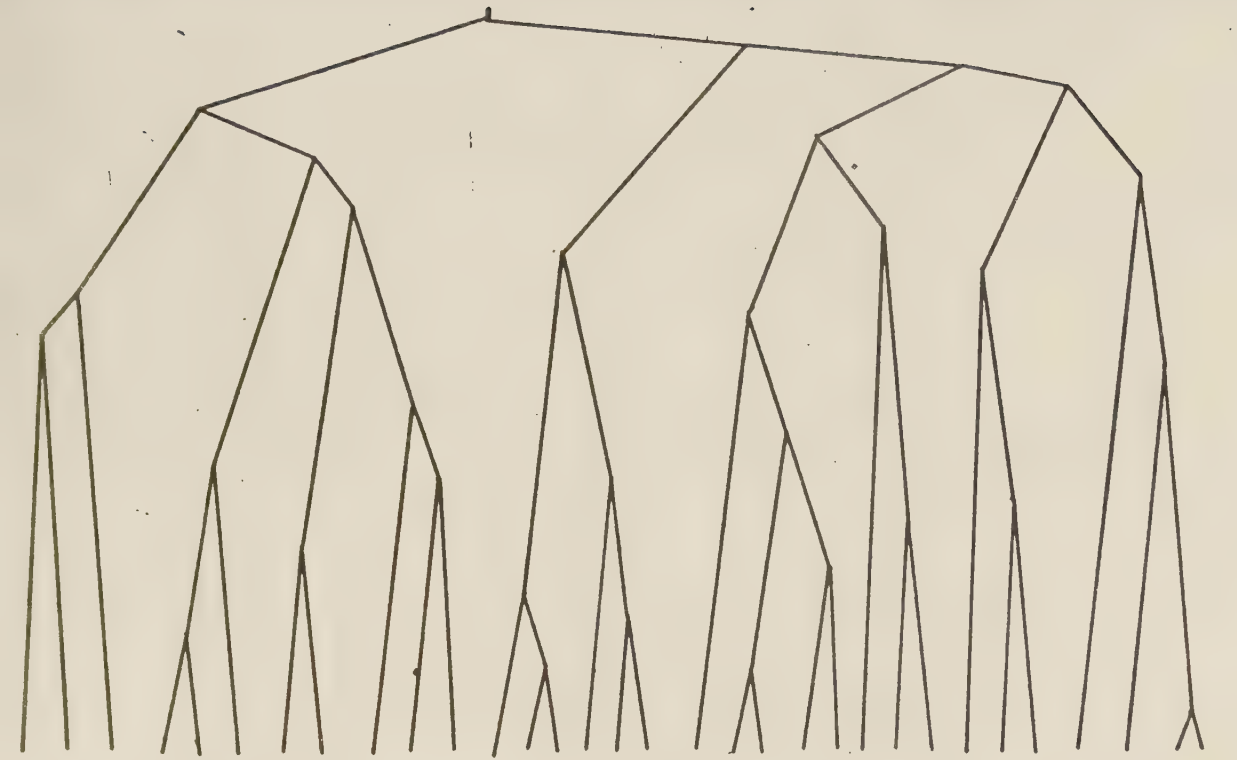



Abb. 13: Dritter Datensatz. Baumgraph n. [illegible]/Ward

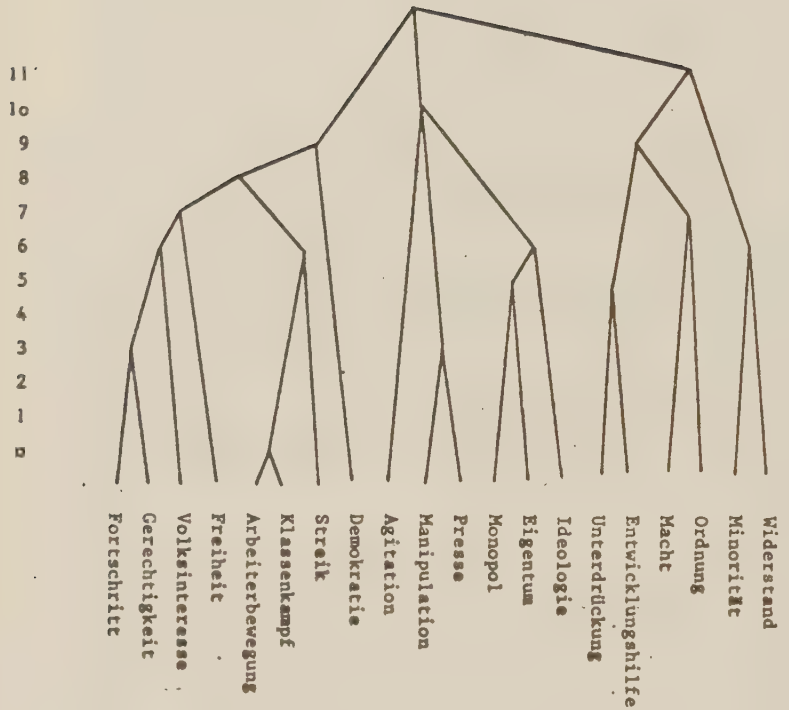



Abb. 14 : Dieser Baumgraph,nach der Maximum-Methode über Item-Distanzen konstruiert,basiert auf den Aufteilungen {2o,21,22,23,25,27,29,3o,31,32} also dem Cluster eines der beiden Hauptäste des Baumes ROSSBACH/Ward des 3.Datensatzes. Er vermittelt damit einen Eindruck von der semantischen Struktur politischer Begriffe bei "linken" Studenten.

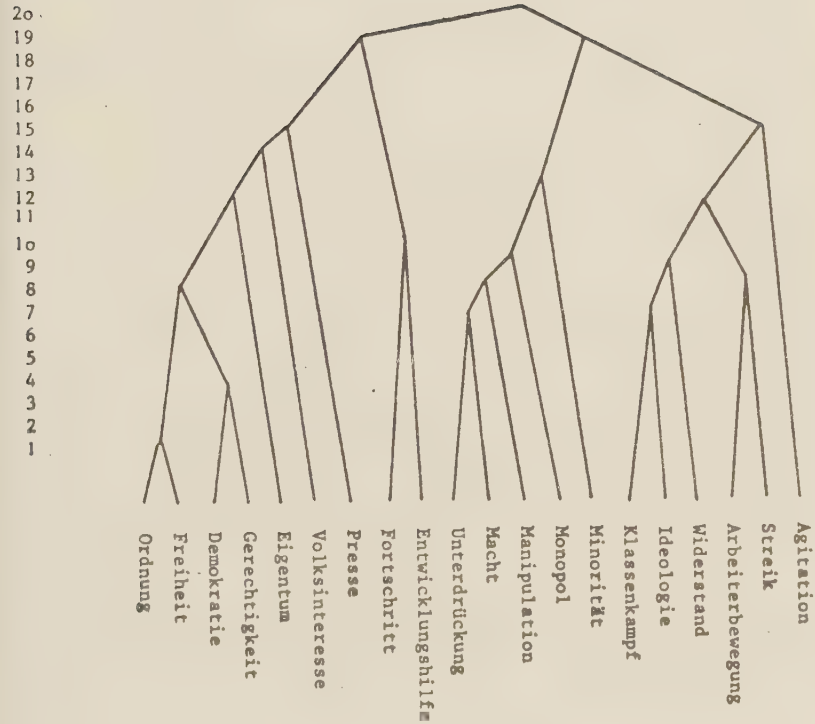



Abb. 15: Dieser Baumgraph, nach der Maximum-Methode über Item-Distanzen konstruiert, basiert auf den Aufteilungen {1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16,17,18,19,24,26}, also dem Cluster eines der beiden Hauptäste des Baumes ROSSBACH/Ward des 3.Datensatzes. Er vermittelt damit einen Eindruck von der semantischen Struktur politischer Begriffe bei "rechten" Studenten.

## 6.4 Zusammenfassung der empirischen Ergebnisse

Schlüsse aus dem Vergleich der drei Datensätze zu ziehen, bedeutet ein Bild ihrer Unterschiede und Übereinstimmung mit Blick auf die Problemstellung zu entwerfen. Die für die Erzeugung der Aufteilungen jeweils bei jedem Datensatz verschiedenen Prinzipien und ihre Bedeutung für die Logik dieses Methodenexperimentes sind in den betreffenden Abschnitten ausführlich behandelt worden. An dieser Stelle sei nur daran erinnert, dass beabsichtigt war, vom ersten über den zweiten zum dritten Datensatz den Schwierigkeitsgrad für die (Re-)Konstruktion der "richtigen" Zweiteilung der Menge der Aufteilungen systematisch zu erhöhen. Dies wurde dadurch erreicht, dass von der vollständigen Kompatibilität der Aufteilungen einer Klasse ausgehend (erster Datensatz) ihre Homogenität laufend abgeschwächt wurde. Was die Ausgangsdaten betrifft, so ist dieses Ziel zweifellos erreicht worden. Das ist an den beiden Extremen, dem ersten und dem dritten Datensatz leicht zu veranschaulichen. Abbildung 16 zeigt die Häufigkeitsverteilung der Distanzen (nach ROSSBACH) bei beiden Datensätzen. Oberhalb der Abszisse sind jeweils die Häufigkeiten der Distanzen zwischen Aufteilungen aus verschiedenen Klassen abgetragen, nach unten die der Innerhalb-Distanzen. Beim ersten Datensatz ist die Überlappung beider Verteilungen (man muss, um sich das klar zu machen, eine von ihnen um die Abszisse gespiegelt denken) deutlich geringer als beim dritten (untere Graphik).

Man darf wohl annehmen, dass das Ausmass der Überlappung in umgekehrtem Verhältnis zur "wahren" Homogenität und Separation der beiden Klassen steht.

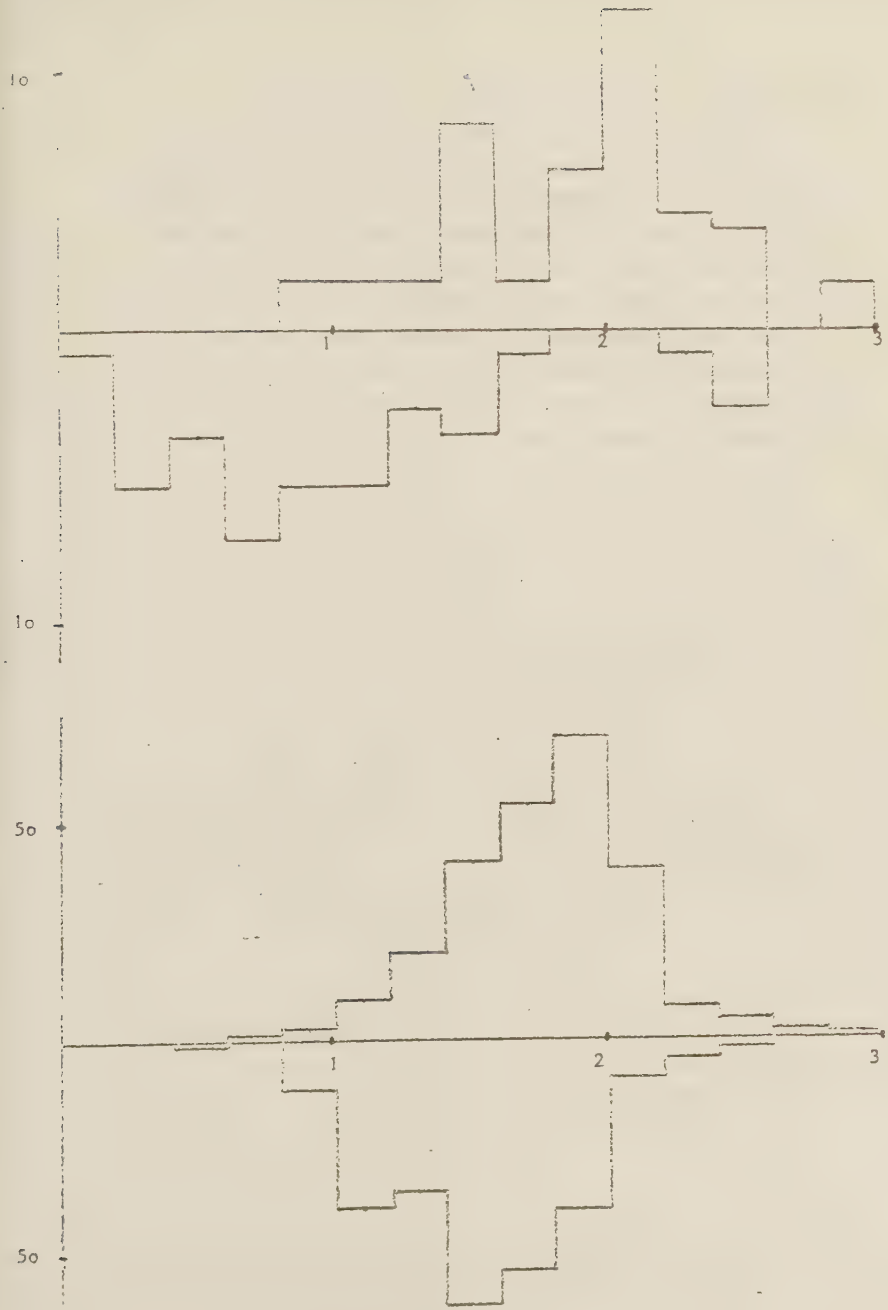



Abb. 16 : Häufigkeitsverteilung der Distanzen (nach ROSSBACH) zwischen den Aufteilungen des 1.(obere Graphik) und des 2.(untere Graphik) Datensatzes. Nach oben sind jeweils die Distanzen zwischen den

Wir schlagen zur Quantifizierung des Überlappungsgrades den folgenden einfachen Index vor:

Der Grundgedanke ist, dass der Grad der Überlappung als umso ausgeprägter angesehen werden muss, je grösser der Anteil der Messwerte der Menge mit dem kleineren Median ist, der sich oberhalb des grösseren Medians befindet - und umgekehrt. Vollständige Überlappung im Sinne dieses Index liegt vor, wenn kein Messwert zwischen den beiden Medianen liegt.

Die Mengen der Messwerte sind

$$X = \{x_1, x_2, x_3, \ldots\ldots, x_i, \ldots, x_s\}$$

und

$$Y = \{y_1, y_2, y_3, \ldots, y_j, \ldots y_t\} \text{ . Mit } Md_x > Md_y$$

$N_x$ ist die Anzahl der Messwerte von X

$N_y$ ist die Anzahl der Messwerte von Y .

Es werden nun die Variablen $P_x$ und $P_y$ definiert als

Definition 28:

$$P_{x_i} = \begin{cases} 1, \text{ wenn } c_i - Md_y < 0 \\ 0, \text{ wenn } x_i - Md_y \geq 0 \end{cases}$$

und

Definition 29:

$$P_{y_i} = \begin{cases} 1, \text{ wenn } y_j - Md_x > 0 \\ 0, \text{ wenn } y_j - Md_y \leq 0 \text{ .} \end{cases}$$

Das Ausmass der Überlappung errechnet sich dann als

Definition 30:

$$\frac{\sum_{i=1}^{s} P_{x_i}}{N_x} + \frac{\sum_{j=1}^{t} P_{y_j}}{N_y}$$

In dieser Weise quantifiziert beträgt die Überlappung zwischen den Verteilungen der Innerhalb- und der Zwischen-Distanzen (ROSSBACH) beim ersten Datensatz nur .14, während sie sich beim dritten auf .49 beläuft.
Derselbe Sachverhalt wird auch durch die Abbildung 17 dargestellt. Dort sind - nun in anderer Form - die Verteilungen der Innerhalb- und der Zwischen-Distanzen aller drei Datensätze repräsentiert. Die durchgezogenen senkrechten Linien kennzeichnen den Bereich einer Standardabweichung oberhalb und unterhalb vom Mittelwert, die gepunktete Linie markiert die Variationsbreite der Werte.

Deutlich nimmt vom ersten über den zweiten zum dritten Datensatz die Distanz zwischen den Mittelwerten der Innerhalb- und der Zwischen-Distanzen eines Datensatzes ab, und die Überlappung ihrer Verteilungen zu. Dass damit das vorgegebene Problem, die - beziehungsweise eine - Zweiteilung in optimal homogene Untergruppen zu finden, schwieriger wird, leuchtet ein. Das geht auch klar aus dem Vergleich der normierten Inhomogenitätswerte (Tabelle 5) hervor: Die Mittelwerte der Datensätze weisen einen Anstieg von .051 beim ersten über .125 beim zweiten bis zu .154 beim dritten auf. Besonders unterstreichen jedoch die grossen Unterschiede zwischen den jeweils optimalen Lösungen (0.0, .092 und .133) die steigende Schwierigkeit der Aufgabe.

Eine andere Frage, die sich bei der gemeinsamen Auswertung der drei Datensätze selbstverständlich stellt, betrifft die Möglichkeit zur Zusammenfassung der Ergebnisse. Dabei wird vor allem die Absicht im Vordergrund stehen, ein möglichst valides Urteil darüber zu erhalten, welches die insgesamt am besten bewährte Methodenkombination ist.

Bevor eine Zusammenfassung der Ergebnisse über alle Datensätze erfolgen kann, muss allerdings zunächst geklärt werden, ob sie auch tatsächlich hinreichend ähnlich sind. Mit der Berechnung des Konkordanzkoeffizienten (nach KENDALL) über die sechs Bewertungen haben wir den Versuch zur Quantifizierung

ihrer Übereinstimmung unternommen. Das Resultat von K = .55 zeigt, dass ein mittelhoher Grad an Übereinstimmung vorliegt, aber wohl nicht genug um beispielsweise die Summation der Rangplätze, die eine Kombination erhalten hat, gerechtfertigt erscheinen zu lassen.

Stattdessen erschien es uns sinnvoller durch Verzicht auf Information das Grundsätzliche deutlicher hervortreten zu lassen.

Zu diesem Zweck wurde unter sorgfältiger Berücksichtigung der Verteilung der Kriteriumswerte innerhalb eines jeweiligen Datensatzes ihre Grobklassifikation nach vier Gütekategorien vorgenommen. Im Sinne des jeweiligen Kriteriums "sehr gute" Zweiteilungen erhielten die Einstufung '++', "gute" '+', schlechte Lösungen '-' und "sehr schlechte" '--'.
Die Gesamtbewertung einer Methodenkombination ergibt sich durch Addition der ihnen insgesamt durch die sechs Tests (je zwei für jeden der drei Datensätze) zuerkannten Plus- und Minuszeichen. (siehe Tabelle 6).

Nach dieser Beurteilung scheint die Verfahrenskombination ROSSBACH/WARD die am Besten bewährte zu sein. Ihr weit überdurchschnittlich gutes Abschneiden steht durchaus in Einklang mit dem, was im übrigen sowohl über das Distanzmass wie auch die Klassifikationsmethode gesagt wurde. Trotzdem wäre es verfehlt zu meinen, dass sich der Wert einer Methodenkombination immer auf einfache Weise aus beiden Bestandteilen ergibt. Vielmehr scheinen in einigen Fällen Wechselwirkungen vorzuliegen. So erzielen Gruppendurchschnitt- und Schwerpunktmethode nur, wenn sie auf ROSSBACH-Distanzen angewendet werden, überraschend gute Ergebnisse, während die allgemein favorisierte Maximum-Methode besonders gut mit den informationstheoretischen Massen harmoniert. Das Distanzmass PAIRBONDS kann nunmehr ohne Bedenken als generell ungeeignet bezeichnet werden.

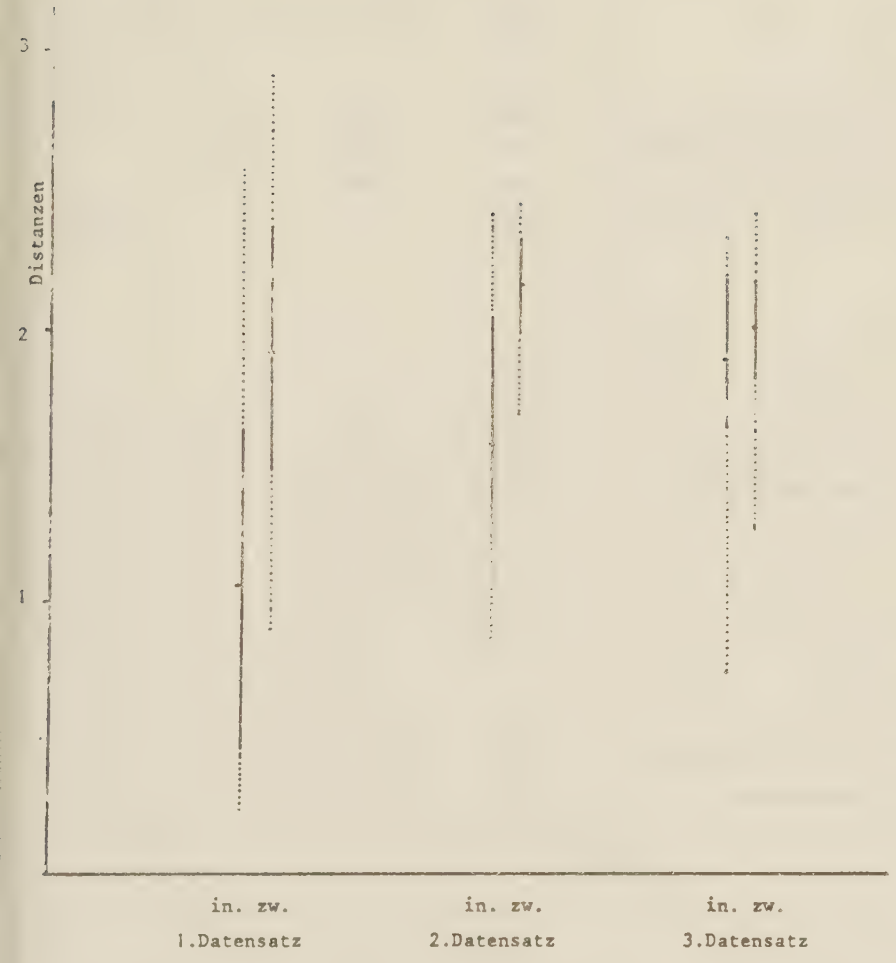



Abb. 17: Die Graphik gibt einen Eindruck von der Verteilung der Distanzen der Aufteilungen der drei Datensätze jeweils innerhalb (in.) und zwischen (zw.) den beiden unterstellten Teilmengen,berechnet nach dem Distanzmaß ROSSBACH (Ordinate). Die Anordnung von links nach rechts auf der Abszisse orientiert sich daran wie ausgeprägt die Zweiteilung ist : Der erste Datensatz ist die "reine" Plasmode,der zweite die mit "noise" und der dritte der empirische Datensatz mit den beiden Gruppen "linke" und "rechte" Studenten. Die durchgezogenen,senkrechten Linien kennzeichnen den Bereich einer Standardabweichung unterhalb und oberhalb vom Mittelwert,die gepunktete Linie markiert die Variationsbreite der Distanzen.Deutlich spiegelt sich die Tendenz zu weniger deutlicher Zweiklassenstruktur in der Zunahme der Überlappung der Verteilungen der Innerhalb- und der Zwischen-Disanzen wider.

| PAIRBONDS | Erster | Zweiter | Dritter | Datensatz |
|---|---|---|---|---|
| Minimum-Methode | .o7o | .139 | - | |
| Maximum-Methode | .o7o | .139 | .171 | |
| Gruppendurchschnitt-Methode | .o7o | .139 | .188 | |
| Schwerpunkt-Methode | .o7o | .139 | .14o | |
| Ward-Methode | .o7o | .139 | .161 | |
| ROSSBACH | | | | |
| Minimum-Methode | .o53 | .12o | - | |
| Maximum-Methode | .o77 | .o92 | .18o | |
| Gruppendurchschnitt-Methode | o.o | .12o | - | |
| Schwerpunkt-Methode | o.o | .o94 | .158 | |
| Ward-Methode | o.o | .1o2 | .133 | |
| INFOTWO | | | | |
| Minimum-Methode | .o53 | .12o | - | |
| Maximum-Methode | .o53 | .o92 | .143 | |
| Gruppendurchschnitt-Methode | .o53 | .12o | .163 | |
| Schwerpunkt-Methode | .o53 | .154 | .14o | |
| Ward-Methode | .o53 | .12o | .141 | |
| INFOTHRY | | | | |
| Minimum-Methode | .o7o | .139 | - | |
| Maximum-Methode | .o53 | .o92 | .146 | |
| Gruppendurchschnitt-Methode | .o53 | .139 | - | |
| Schwerpunkt-Methode | .o53 | .16o | .154 | |
| Ward-Methode | .o53 | .139 | .144 | |
| ARITHMET. MITTEL | .o51 | .125 | .154 | |
| STREUUNG | .o25 | .o22 | .o17 | |

Tab. 5 : Bewertung der Ergebnisse durch das normierte Inhomogenitätsmaß

| PAIRBONDS | 1.Datensatz Inh. | 1.Datensatz g.o.f. | 2.Datensatz Inh. | 2.Datensatz g.o.f. | 3.Datensatz Inh. | 3.Datensatz g.o.f | Anzahl '+' | Anzahl '-' |
|---|---|---|---|---|---|---|---|---|
| Minimum-M. | -- | -- | -- | -- | -- | -- | o | 12 |
| Maximum-M. | -- | -- | -- | -- | - | -- | o | 11 |
| Gruppend.-M. | -- | -- | -- | -- | -- | -- | o | 12 |
| Schwerp.-M. | - | -- | -- | -- | ++ | + | 3 | 7 |
| Ward-M. | - | -- | -- | -- | -- | - | o | 1o |
| ROSSBACH | | | | | | | | |
| Minimum-M. | -- | + | + | + | -- | -- | 3 | 6 |
| Maximum-M. | -- | + | ++ | ++ | -- | - | 5 | 5 |
| Gruppend.-M. | ++ | ++ | + | + | -- | - | 6 | 3 |
| Schwerp.-M. | ++ | ++ | ++ | ++ | - | + | 9 | 1 |
| Ward-M. | ++ | ++ | ++ | ++ | ++ | + | 11 | o |
| INFOTWO | | | | | | | | |
| Minimum-M. | - | -- | + | - | -- | -- | 1 | 9 |
| Maximum-M. | - | - | ++ | + | + | + | 6 | 2 |
| Gruppend.-M. | - | - | + | -- | -- | - | 1 | 7 |
| Schwerp.-M. | - | - | -- | - | ++ | + | 3 | 5 |
| Ward-M. | - | - | + | - | ++ | + | 4 | 3 |
| INFOTHRY | | | | | | | | |
| Minimum-M. | -- | -- | -- | -- | -- | -- | o | 12 |
| Maximum-M. | - | - | ++ | + | + | ++ | 6 | 2 |
| Gruppend.-M. | - | -- | -- | -- | -- | -- | o | 11 |
| Schwerp.-M. | - | - | -- | -- | - | -- | o | 9 |
| Ward-M. | - | - | -- | -- | + | + | 2 | 6 |

Tab. 6 : Zusammenfassung der Bewertungen für alle drei Datensätze durch Summation der '+' und '-' ,nach denen sie grob klassifiziert worden sind. '++' ='sehr gut','+' = 'gut','-' = 'schlecht','--' = 'sehr schlecht'.

## 7 DISKUSSION UND ZUSAMMENFASSUNG

Die vorliegende Arbeit hat sich um Ziel gesetzt, auf eine bislang unerkannte Möglichkeit der Datenanalyse aufmerksam zu machen und gleichzeitig die noch notwendigen Vorarbeiten zu ihrer praktischen Durchführung zu leisten.

Der von uns "inverse Clusteranalyse" genannte Ansatz, soll vor allem dazu dienen, das tatsächliche Vorliegen der Voraussetzungen der normalen Hierarchischen Clusteranalyse von Sorting-Daten abzuklären beziehungsweise durch geeignete Gruppenbildung erst herzustellen. Die besondere Bedeutung dieser Vorgehendsweise liegt in der darin schon angedeuteten differentiellen Analyse von Aufteilungen.

Da Aufteilungen eine ausserordentlich leicht erhebbare Form von Daten sind, könnte diese methodische Ergänzung unter Umständen ein neues und weites Feld von Anwendungen erschliessen.
Das Ziel der "inversen Clusteranalyse" besteht in der Zusammenfassung von Aufteilungen zu strukturell homogenen Gruppen. Ein hohes Mass an Homogenität ist die notwendige - und bisher stets im Sinne eines allgemeinpsychologischen Ansatzes als gegeben unterstellte - Voraussetzung für die sinnvolle Zusammenfassung aller Sortings zur Berechnung der quantitativen Beziehungen (Ähnlichkeiten, Distanzen) zwischen den Paaren der sortierten Items.

Zu diesem Zweck musste ein geeignetes Distanzmass für Aufteilungen gefunden werden. Vier solcher Masse wurden schliesslich zur ausführlichen Diskussion und praktischen Erprobung ausgewählt. Sowohl bezüglich seiner theoretsichen Voraussetzungen als auch bei der empirischen Testung erwies sich das vom Verfasser für diesen Zweck entwickelte Distanzmass (in Kombination mit der Hierarchischen Klassifikationsmethode von WARD) als insgesamt am geeignetsten.

Dieses Urteil hat jedoch notwendig nur vorläufigen Charakter. Es kann sich weder auf einen vollständigen Vergleich aller bekannten in Frage kommenden Masse stützen, noch ist überhaupt die Entwicklung derartiger Indizes als abgeschlossen anzusehen. Zum mindesten liegt aber mit dieser Arbeit ein ausgefeiltes Instrumentarium vor. Es steht jederzeit zur methodenkritischen Erprobung weiterer Distanzmasse zur Verfügung, von deren Verwendung im Rahmen der inversen Clusteranalyse man sich etwas verspricht.

LITERATUR

ABRAMOWITZ, M. and STEGUN, I.A. : Handbook of mathematical functions with formulas, graphs and mathematical tables.
U.S. Govt. Printing Office, Washington, D.C. (1968).

ANDERBERG, M. R.: Cluster analysis für applications.
Academic Press, New York (1973).

ARABIE, P. and BOORMAN, S. A. : Structural measures and the method of sorting.
In: SHEPARD, R.N. et al. (eds.) Multidimensional scaling, Vol. 1 New York, Seminar Press (1972).

ARABIE, P. and BOORMAN, S. A.: Multidimensional scaling of measures of distance between particians,
Journal Math. Psych., 10, 148-203 (1973).

BOORMAN, S. A. and ARABIE, P. : Structural measures and the method of sorting.
In: SHEPARD, R.N. et al. (eds.) Multidimensional scaling, Vol. 1, New York, Seminar Press (1972).

CATTELL, R. B. and SULLIVAN, W. : The scientific nature of factors: A demonstration by cups of coffee
Behav. Sci., 7, 184-193 (1962).

CHOMSKY, N. and MILLER, G. A.: Indtroduction to the formal analysis of natural languages.
In: LUCE, R.D. et al. (eds.) Handbook of Mathematical Psychology, Vol. 2, Wiley, New York (1963).

COLE, A. J. and WISHART, D. : An improved algorithm for the Jardine-Sibson method of generating overlapping clusters,
Comp. J., 13, 156-163 (1970).

EDWARDS, A.W. F. and CAVALLI-SFORZA, L. L. : A method for cluster analysis. Biometrics, 21, No. 2, 362-375 (1965)

FRIEDELL, M. F. : Organizations as semilattices.
Amer. Sociol. Rev., 32, 46-54 (1967).

GALANTER, E. : An axiomatic and ecperimental study of sensory and measure.
Psych. Rev., 63, 16-28 (1956).

GOETHE, J. W. : Erster Entwurf zur allgemeinen Einleitung in die vergleichende Anatomie. (1796)

GOODMAN, N. : Fact, fiction and forecast.
Bobbs-Merill, Indianapolis (1955).

GOWER, J. A. : A comparison of some methods of cluster analysis.
Biometrics, 23, 623-637 (1967).

HAWKINS, D. : Taxonomy and information.
In: R.S. Cohen and M. Wartwsky (eds.)
Boston studies in the philosophy of science. Vol.3, 41
Humanities Press, New York (1968).

HORKHEIMER, M. und ADORNO, Th. W. : Dialektik der Aufklärung.
Querido, Amsterdam (1944).

HUBERT, L. : Some extensions of Johnson's hierarchical cluster algorithms.
Psychometrica, 37, 261-274 (1972).

JARDINE, N. and SIBSON, R. : The construction of hierarchic and non-hierarchic classifications.
Computer Hournal, 11, 117-184 (1968).

JARDINE, N. and SIBSON, R. : Mathematical taxonomy.
Wiley, New York (1971).

JOHNSON, S. C. : Hierarchical clustering schemes.
Psychometrica, 32, 241-254 (1967).

KATZ, J. J. and FODOR, J. A. : The structure of a semantic theory.
In: Fodor and Katz (eds.). The structure of language.
Prentice Hall, Englewood Cliffs, New York (1964).

LANCE, G. N. and WILLIAMS, W. T. : A general theory of classificatory sorting strategies: 1. Hierarchical systems.
Computer Journal, 9, 60-64 (1967).

LING, R. F. : A probability theory of cluster analysis.
J. Amer. Statist. Assoc., 68, 159-164 (1973).

LORENZ, K. : Gestaltwahrnehmung als Quelle wissenschaftlicher Erkenntnis.
Zeitschrift experimentelle angewandte Psychologie, 4, 121-143 (1959).

LORENZ, K. : Die Rückseite des Spielgels.
München: Piper (1973).

MILLER, G. A. : A psychological method to investigate verbal concepts.
J. Math. Psych., 6, 169-191 (1969).

NAEF, A. : Idealistische Morphologie und Genetik.
O. Fischer, Jena (1919).

NAEF, A. : Die Gestalt als Begriff und Idee.
In: Handbuch der Vergleichenden Anatomie der Wirbeltiere,
Urban und Schwarzenberg, Berlin (1931)

POPPER, K. : Logik der Forschung.
Mohr, Tübingen (1966)

RASKI, C. : A metric space of discrete probability distributions.
Information and Control, 4, 371-377 (1961)

RATTINGER, H. : Eine einfache Methode und ein FORTRAN-Programm zur Ermittlung von Cliquen.
Zeitschrift Sozialpsychologie, 4, 5-14 (1973)

REMANE, A.: Die Grundlagen des natürlichen Systems der Vergleichenden Anatomie und der Phylogenetik.
Akademische Verlagsgesellschaft, Leipzig (1956)

RESTLE, F. : A metric and an ordering of sets.
Psychometrika, 24, 207-220 (1959)

RESTLE, F.: Psychology of judgement and choice.
Wiley, New York (1961)

ROSSBACH, H. : Erprobung von empirischen Verfahren zur Analyse der semantischen Struktur lexikalischer Begriffe.
Diplomarbeit, Hamburg (1969)

SOKAL, R. R. and SNEATH, P.H. A.: Principles of numerical taxonomy.
Freeman, San Francisco, Calif. (1963)

SZASZ, G.: Introduction to lattice theory.
Academic Press, New York (1963)

THURSTONE, L. L. : Multiple factor analysis.
Univ. Chicago Press, Chicago (1957)

ÜBERLA, K. : Faktorenanalyse.
Springer, Heidelberg (1968)

WALLACE, D. L. : Clustering.
In: D.L. Sills (ed.) International Encyclopedia of the Social Sciences, Vol. 2, 519-524,
Macmillen and Free Press (1968)

WARD, H. J. : Hierarchical grouping to optimize an objective function.
J. Amer. Statist. Assoc., 58, 236-244 (1963)

WATANABE, S. : Knowing and guessing.
Wiley, New York (1969)

WISHART, D. : An algorithm for hierarchical classifications.
Biometrics, 22, 165-170 (1969)

# ANHANG

a) Die Aufteilungen des dritten Datensatzes
(Tabelle A)

b) Die Matrizen der Distanzen nach PAIRBONDS,ROSSBACH,
INFOTWO und INFOTHRY für alle drei Datensätze
(Tabelle B - M)

c) Beispiel eines Computer-Ausdrucks für Baumgraphen
(Abbildung A)

R1 = {(6),(12),(17,18),(5,15),4,7,10,11,13,14,20),(1,2,3,8,9,16,19),}

R2 = {(4,10,11,14,17),(2,3,9,8,12,16,18),(1,5,6,7,13,15,19,20)}

R3 = {(1,12,14,18),(2,3,9,16),(4,7,10,11,17),(8,9),(5,6,13,15,20)}

R4 = {(6,17,18),(7,10,11,14),(3,8,9,16,19),(1,12),(4),(2),(5,13,15,20)}

R5 = {(6,13,20),(1,12),(3,8,9,16,19),(4,11,14),(2,5,15),(7,10,17,18)}

R6 = {(2,5,15,19),(6,17,18),(1,12),(1o,13,20),(4,7,8,11,14,16),(3,9)}

R7 = {(5,15),(1,12,16),(3,8,9,18,19),(2,4,6,7,10,11,13,14,17,20)}

R8 = {(6,17,20),(2,12),(4,7,10),(5,11,13,14,15),(1,3,8,9,16,18,19)}

R9 = {(20),(1,12),(13,6),(10,7,17,11,15,14,5),(2,9,8,18,19,16,4,3)}

R10 = {(2,3,8,9,16),(5),(1),(18),(4,7,19),(20),(11,13,14,15),(12),(1o)(6,17

R11 = {(1,2,3,5,8,9,12,15,16,18,19),(4,6,7,11,13,14,17,20)}

R12 = {(2,4,5,13,14,15,20),(18,17),(9,16,3,19,8),(6),(12),(1,7,10,11)}

R13 = {(7,11,13,14,17,20),(3,8,9,16),(1,12),(19),(1o),(4),(6)}

R14 = {(1,2,7,8,9,15,16),(12,19),(5,6,13,14,20),(3,4,11,17),(10,18)}

R15 = {(4,7,10,11,15,17,18,20),(1,2,12,16),(3,8,9,19),(5,6,13,14)}

R16 = {(2,8,9,19),(4,6,13,18),(5,14,15,20),(1,3,12,16),(7,10,11,17)}

R17 = {(2,5,13,14,15),(1,3,9,16,20),(12,18),(8),(4,6,7,11,10,17,19)}

R18 = {(1,2,3,9,16),(7,17),(8,10,19),(5,6,11,13,14,15,20),(4,12),(18)}

R19 = {(7,8,11,10,12,18,19),(2,3,9,16),(4,7),(1,5,6,13,14,15,20)}

R20 = {(17,18),(4,7,8,10,11,12,14,19),(3,15,5,9,2,6,16,20,13,1)}

R21 = {(4),(6),(16),(17,18,20),(8,7,10,19),(11,12),(3,5,2,1,9,13,14,15)}

R22 = {(12),(8,14),(1,2,3,9,16),(6,17,18),(5,13,15),(4,7,Lo,11,19,20)}

R23 = {(1,2,3,4,5,9,13,14,15,16),(18),(6,7,8,10,11,12,17,19,20)}

R24 = {(7,10,11,17),(4),(12),(1),(2,3,8,9,16,18,19),(5,6,13,14,15,20)}

R25 = {(1,2,3,5,9,12,13,15,16,19),(10,11,6,17,20,7),(8),(14),(4),(18)}

R26 = {(12),(8),(1o,19),(4,7,11,17,18),(2,3,9,16),(1,5,6,13,14,15,2o)}

R27 = {(3,9,16),(7,11),(13,15),(6,8,17,18),(5,14),(1,2),(4,19,20),(10,12)}

R28 = {(7,10),(1,2,3,16),(9,17,18,20),(4,11,12,14),(5,13,15),(8),(6),(19)}

R29 = {(6),(4),(7,8,10,11,12,14,17,18,19,20),(1,2,3,5,9,13,15,16)}

R30 = {(1,16,19),(3,10,17,18),(8,11,12),(2,9,13,15,20),(4,5,6,7,14)}

R31 = {(10,11,19,20),(12),(1,9,13,15,16),(4,5,6,14),(2),(3,8,17,18,7)}

R32 = {(3,8,10,17,18 19),(20),(7,11,12),(1,2,4,5,6,9,13,14,15,16)}

Tabelle A : 3. Datensatz. Die Aufteilungen der 20 politischen Begriffe durch in linken politischen Verbindungen organisierte (R1-R16) und Verbindungsstudenten (R17-R32).

Tab. B : Distanzen zwischen den Aufteilungen des ersten Datensatzes berechnet nach PAIRBONDS.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 2 | | | | | | | | | | | | |
| 3 | 22 | 20 | | | | | | | | | | | |
| 4 | 33 | 31 | 11 | | | | | | | | | | |
| 5 | 35 | 33 | 13 | 2 | | | | | | | | | |
| 6 | 37 | 35 | 15 | 4 | 2 | | | | | | | | |
| 7 | 38 | 36 | 16 | 5 | 3 | 1 | | | | | | | |
| 8 | 35 | 35 | 25 | 26 | 24 | 22 | 21 | | | | | | |
| 9 | 39 | 39 | 25 | 16 | 14 | 12 | 11 | 10 | | | | | |
| 10 | 38 | 36 | 22 | 13 | 11 | 9 | 8 | 13 | 3 | | | | |
| 11 | 37 | 35 | 21 | 10 | 8 | 6 | 5 | 16 | 6 | 3 | | | |
| 12 | 35 | 33 | 19 | 8 | 6 | 4 | 3 | 18 | 8 | 5 | 2 | | |
| 13 | 37 | 35 | 19 | 8 | 6 | 4 | 3 | 20 | 10 | 7 | 4 | 2 | |
| 14 | 38 | 36 | 18 | 7 | 5 | 3 | 2 | 21 | 11 | 8 | 5 | 3 | 1 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. C : Distanzen zwischen den Aufteilungen des ersten Datensatzes berechnet nach ROSSBACH,

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 0,736 | | | | | | | | | | | | |
| 3 | 1.671 | 0.686 | | | | | | | | | | | |
| 4 | 2.383 | 1.433 | 0,917 | | | | | | | | | | |
| 5 | 2.453 | 1.477 | 1,010 | 0.366 | | | | | | | | | |
| 6 | 2.523 | 1.519 | 1.145 | 0.729 | 0,365 | | | | | | | | |
| 7 | 2.557 | 1.540 | 1.213 | 0.910 | 0.546 | 0.515 | | | | | | | |
| 8 | 2.258 | 2.135 | 1.811 | 2.275 | 1.885 | 1.829 | 1.301 | | | | | | |
| 9 | 2.583 | 2.265 | 2.193 | 1.998 | 1.622 | 1.579 | 1.058 | 0.719 | | | | | |
| 10 | 2.902 | 2.074 | 2.117 | 1.943 | 1.571 | 1.531 | 1.012 | 0.928 | 0.219 | | | | |
| 11 | 2.806 | 1.897 | 2.169 | 1.833 | 1.464 | 1.428 | 0.910 | 1.033 | 0.600 | 0.383 | | | |
| 12 | 2.453 | 1.477 | 2.119 | 1.797 | 1.430 | 1.396 | 0.879 | 1.168 | 0.740 | 0.636 | 0.366 | | |
| 13 | 2.523 | 1.519 | 2.306 | 2.127 | 2.095 | 2.063 | 2.046 | 1.235 | 0.990 | 0.887 | 0.729 | 0.365 | |
| 14 | 2.557 | 1.540 | 2.282 | 2.109 | 2.078 | 2.046 | 2.031 | 1.268 | 1.114 | 1.012 | 0.910 | 0.546 | 0.515 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. D: Distanzen zwischen den Aufteilungen des ersten Datensatzes berechnet nach INFOTWO.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 0.230 | | | | | | | | | | | | |
| 3 | 0.973 | 0.743 | | | | | | | | | | | |
| 4 | 1.981 | 1.752 | 1.008 | | | | | | | | | | |
| 5 | 2.211 | 1.981 | 1.238 | 0.230 | | | | | | | | | |
| 6 | 2.441 | 2.211 | 1.468 | 0.459 | 0.230 | | | | | | | | |
| 7 | 2.607 | 2.378 | 1.634 | 0.626 | 0.396 | 0.167 | | | | | | | |
| 8 | 1.965 | 2.068 | 2.260 | 2.585 | 2.356 | 2.126 | 1.959 | | | | | | |
| 9 | 2.441 | 2.544 | 2.593 | 1.918 | 1.689 | 1.459 | 1.293 | 0.667 | | | | | |
| 10 | 2.503 | 2.274 | 2.323 | 1.648 | 1.418 | 1.189 | 1.022 | 0.937 | 0.270 | | | | |
| 11 | 2.441 | 2.211 | 2.260 | 1.252 | 1.022 | 0.793 | 0.626 | 1.333 | 0.667 | 0.396 | | | |
| 12 | 2.211 | 1.981 | 2.031 | 1.022 | 0.793 | 0.563 | 0.396 | 1.563 | 0.896 | 0.626 | 0.230 | | |
| 13 | 2.441 | 2.211 | 2.134 | 1.126 | 0.896 | 0.667 | 0.500 | 1.793 | 1.126 | 0.855 | 0.459 | 0.230 | |
| 14 | 2.607 | 2.378 | 1.968 | 0.959 | 0.730 | 0.500 | 0.333 | 1.959 | 1.293 | 1.022 | 0.626 | 0.396 | 0.167 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. 2 : Distanzen zwischen den Aufteilungen des ersten Datensatzes berechnet nach INFOTHRY.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 0.221 | | | | | | | | | | | | |
| 3 | 0.545 | 0.417 | | | | | | | | | | | |
| 4 | 0.709 | 0.627 | 0.361 | | | | | | | | | | |
| 5 | 0.732 | 0.656 | 0.410 | 0.076 | | | | | | | | | |
| 6 | 0.751 | 0.680 | 0.451 | 0.141 | 0.071 | | | | | | | | |
| 7 | 0.763 | 0.696 | 0.478 | 0.183 | 0.116 | 0.049 | | | | | | | |
| 8 | 0.928 | 0.906 | 0.821 | 0.756 | 0.689 | 0.622 | 0.573 | | | | | | |
| 9 | 0.908 | 0.891 | 0.798 | 0.561 | 0.494 | 0.427 | 0.378 | 0.314 | | | | | |
| 10 | 0.877 | 0.796 | 0.714 | 0.482 | 0.415 | 0.348 | 0.299 | 0.391 | 0.113 | | | | |
| 11 | 0.808 | 0.732 | 0.661 | 0.366 | 0.299 | 0.232 | 0.183 | 0.477 | 0.239 | 0.142 | | | |
| 12 | 0.732 | 0.656 | 0.594 | 0.299 | 0.232 | 0.165 | 0.116 | 0.517 | 0.297 | 0.207 | 0.076 | | |
| 13 | 0.751 | 0.680 | 0.595 | 0.314 | 0.250 | 0.186 | 0.139 | 0.551 | 0.346 | 0.263 | 0.141 | 0.071 | |
| 14 | 0.763 | 0.696 | 0.549 | 0.268 | 0.204 | 0.139 | 0.093 | 0.573 | 0.378 | 0.299 | 0.183 | 0.116 | 0.049 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 16 | | | | | | | | | | | | |
| 3 | 23 | 13 | | | | | | | | | | | |
| 4 | 25 | 13 | 14 | | | | | | | | | | |
| 5 | 21 | 11 | 8 | 8 | | | | | | | | | |
| 6 | 19 | 11 | 10 | 6 | 4 | | | | | | | | |
| 7 | 26 | 18 | 15 | 17 | 13 | 13 | | | | | | | |
| 8 | 29 | 29 | 26 | 26 | 24 | 22 | 29 | | | | | | |
| 9 | 29 | 25 | 22 | 20 | 18 | 18 | 27 | 26 | | | | | |
| 10 | 30 | 24 | 23 | 21 | 19 | 19 | 24 | 21 | 21 | | | | |
| 11 | 26 | 18 | 17 | 13 | 13 | 11 | 22 | 23 | 11 | 18 | | | |
| 12 | 21 | 15 | 12 | 8 | 6 | 6 | 17 | 24 | 14 | 19 | 9 | | |
| 13 | 23 | 13 | 12 | 6 | 6 | 4 | 15 | 22 | 14 | 17 | 7 | 6 | |
| 14 | 36 | 32 | 33 | 31 | 33 | 31 | 34 | 17 | 33 | 30 | 32 | 31 | 29 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. F : Distanzen zwischen den Aufteilungen des zweiten
Integrates berechnet nach RAYBOULD

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 1.286 | | | | | | | | | | | | |
| 3 | 2.026 | 1.581 | | | | | | | | | | | |
| 4 | 2.428 | 1.883 | 2.233 | | | | | | | | | | |
| 5 | 1.618 | 1.015 | 1.297 | 2.129 | | | | | | | | | |
| 6 | 1.230 | 1.424 | 1.547 | 2.095 | 1.396 | | | | | | | | |
| 7 | 1.837 | 1.524 | 1.302 | 2.285 | 1.423 | 1.665 | | | | | | | |
| 8 | 2.163 | 2.407 | 2.176 | 2.446 | 2.218 | 1.829 | 2.277 | | | | | | |
| 9 | 2.184 | 2.395 | 2.099 | 2.338 | 2.024 | 2.290 | 2.448 | 1.558 | | | | | |
| 10 | 2.047 | 2.255 | 2.198 | 2.356 | 2.119 | 2.306 | 1.971 | 1.169 | 1.546 | | | | |
| 11 | 2.172 | 2.113 | 2.079 | 2.215 | 2.215 | 2.176 | 2.398 | 1.446 | 1.220 | 1.780 | | | |
| 12 | 1.784 | 2.250 | 1.838 | 2.129 | 1.680 | 2.095 | 2.285 | 2.135 | 1.302 | 2.036 | 1.481 | | |
| 13 | 2.373 | 2.298 | 2.192 | 2.095 | 2.095 | 2.063 | 2.241 | 1.829 | 1.074 | 1.737 | 0.859 | 2.095 | |
| 14 | 2.024 | 2.073 | 1.878 | 1.989 | 2.385 | 1.979 | 2.255 | 1.142 | 2.068 | 2.084 | 1.962 | 1.989 | 1.417 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. G : Distanzen zwischen den Aufteilungen des zweiten Datensatzes berechnet nach ROSSBACH.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 1.498 | | | | | | | | | | | | |
| 3 | 2.206 | 1.500 | | | | | | | | | | | |
| 4 | 2.768 | 1.730 | 1.896 | | | | | | | | | | |
| 5 | 1.998 | 1.293 | 1.000 | 1.230 | | | | | | | | | |
| 6 | 1.768 | 1.396 | 1.230 | 1.000 | 0.563 | | | | | | | | |
| 7 | 2.394 | 2.022 | 1.522 | 2.293 | 1.522 | 1.626 | | | | | | | |
| 8 | 2.309 | 2.855 | 2.356 | 2.793 | 2.356 | 2.126 | 2.752 | | | | | | |
| 9 | 2.528 | 2.823 | 2.323 | 2.427 | 1.990 | 2.093 | 3.053 | 2.093 | | | | | |
| 10 | 2.569 | 2.531 | 2.364 | 2.468 | 2.031 | 2.134 | 2.427 | 1.593 | 1.768 | | | | |
| 11 | 2.498 | 2.126 | 1.959 | 1.730 | 1.626 | 1.396 | 2.689 | 2.063 | 1.156 | 1.738 | | | |
| 12 | 2.102 | 2.063 | 1.563 | 1.333 | 0.896 | 1.000 | 2.293 | 2.459 | 1.427 | 2.134 | 1.063 | | |
| 13 | 2.435 | 1.730 | 1.563 | 1.000 | 0.896 | 0.667 | 1.959 | 2.126 | 1.427 | 1.801 | 0.730 | 1.000 | |
| 14 | 2.246 | 2.333 | 2.293 | 2.396 | 2.626 | 2.396 | 2.563 | 1.145 | 2.238 | 2.071 | 2.333 | 2.396 | 2.063 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. H : Distanzen zwischen den Aufteilungen des zweiten Datensatzes berechnet nach INFOTWO.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | 0,579 | | | | | | | | | | | | |
| 3 | 0,730 | 0,496 | | | | | | | | | | | |
| 4 | 0,772 | 0,505 | 0,529 | | | | | | | | | | |
| 5 | 0,514 | 0,397 | 0,314 | 0,343 | | | | | | | | | |
| 6 | 0,564 | 0,404 | 0,361 | 0,279 | 0,165 | | | | | | | | |
| 7 | 0,820 | 0,595 | 0,522 | 0,639 | 0,468 | 0,476 | | | | | | | |
| 8 | 0,879 | 0,678 | 0,764 | 0,779 | 0,689 | 0,622 | 0,892 | | | | | | |
| 9 | 0,856 | 0,825 | 0,714 | 0,677 | 0,582 | 0,584 | 0,893 | 0,779 | | | | | |
| 10 | 0,833 | 0,778 | 0,727 | 0,668 | 0,594 | 0,595 | 0,787 | 0,659 | 0,658 | | | | |
| 11 | 0,763 | 0,722 | 0,5 | 0,482 | 0,454 | 0,389 | 0,750 | 0,683 | 0,420 | 0,575 | | | |
| 12 | 0,549 | 0,575 | 0,457 | 0,372 | 0,262 | 0,279 | 0,639 | 0,719 | 0,462 | 0,624 | 0,327 | | |
| 13 | 0,579 | 0,632 | 0,493 | 0,279 | 0,250 | 0,186 | 0,546 | 0,622 | 0,439 | 0,527 | 0,224 | 0,279 | |
| 14 | 0,711 | 0,717 | 0,795 | 0,737 | 0,768 | 0,701 | 0,898 | 0,604 | 0,852 | 0,821 | 0,772 | 0,737 | 0,634 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Tab. I: Distanzen zwischen den Aufteilungen des zweiten Datensatzes berechnet nach INFOTHRY.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 61 | | | | | | | | | | | | | | | | | | | | |
| 3 | 47 | 46 | | | | | | | | | | | | | | | | | | | |
| 4 | 32 | 55 | 31 | | | | | | | | | | | | | | | | | | |
| 5 | 36 | 57 | 33 | 20 | | | | | | | | | | | | | | | | | |
| 6 | 45 | 70 | 48 | 33 | 33 | | | | | | | | | | | | | | | | |
| 7 | 45 | 70 | 58 | 53 | 51 | 62 | | | | | | | | | | | | | | | |
| 8 | 34 | 61 | 47 | 32 | 36 | 55 | 55 | | | | | | | | | | | | | | |
| 9 | 51 | 60 | 52 | 41 | 43 | 60 | 62 | 45 | | | | | | | | | | | | | |
| 10 | 35 | 53 | 37 | 28 | 32 | 39 | 63 | 30 | 43 | | | | | | | | | | | | |
| 11 | 49 | 64 | 76 | 77 | 71 | 82 | 50 | 65 | 74 | 81 | | | | | | | | | | | |
| 12 | 40 | 67 | 45 | 24 | 30 | 49 | 57 | 42 | 55 | 40 | 85 | | | | | | | | | | |
| 13 | 38 | 59 | 43 | 33 | 28 | 37 | 47 | 44 | 49 | 30 | 63 | 40 | | | | | | | | | |
| 14 | 55 | 62 | 48 | 53 | 47 | 56 | 74 | 57 | 68 | 45 | 80 | 55 | 51 | | | | | | | | |
| 15 | 42 | 69 | 43 | 46 | 42 | 61 | 51 | 52 | 55 | 52 | 75 | 57 | 50 | 55 | | | | | | | |
| 16 | 44 | 63 | 35 | 32 | 36 | 49 | 55 | 48 | 43 | 40 | 75 | 40 | 42 | 41 | [illegible] | | | | | | |
| 17 | 53 | 67 | 41 | 48 | 48 | 55 | 59 | 48 | 67 | 42 | 81 | 48 | 50 | 61 | 58 | 40 | | | | | |
| 18 | 44 | 61 | 33 | 40 | 42 | 57 | 65 | 38 | 57 | 32 | 81 | 46 | 42 | 43 | 52 | 40 | 48 | | | | |
| 19 | 65 | 54 | 36 | 49 | 53 | 66 | 78 | 59 | 68 | 49 | 90 | 53 | 61 | 57 | 61 | 54 | 55 | 37 | | | |
| 20 | 70 | 75 | 59 | 66 | 70 | 71 | 93 | 82 | 91 | 72 | 91 | 74 | 82 | 75 | 88 | 80 | 64 | 60 | 51 | | |
| 21 | 60 | 77 | 53 | 50 | 50 | 55 | 81 | 52 | 73 | 44 | 91 | 48 | 54 | 57 | 64 | 54 | 48 | 44 | 47 | 56 | |
| 22 | 32 | 61 | 35 | 34 | 44 | 41 | 63 | 44 | 61 | 30 | 79 | 48 | 45 | 29 | 46 | 44 | 34 | 29 | 41 | [illegible] | [illegible] |
| 23 | 83 | 94 | 78 | 85 | 83 | 86 | 90 | 85 | 92 | 79 | 88 | 75 | 83 | 84 | 89 | 81 | 61 | 71 | 56 | 63 | [illegible] |
| 24 | 42 | 45 | 29 | 30 | 34 | 59 | 55 | 34 | 31 | 36 | 65 | 34 | 42 | 47 | 48 | 32 | 36 | 32 | 37 | 36 | [illegible] |
| 25 | 50 | 67 | 59 | 58 | 58 | 67 | 73 | 62 | 75 | 64 | 49 | 68 | 62 | 71 | 66 | 38 | 58 | 62 | 76 | 60 | 57 |
| 26 | 54 | 49 | 25 | 42 | 42 | 53 | 69 | 54 | 61 | 38 | 87 | 46 | 44 | 43 | 46 | 52 | 40 | 48 | 23 | 60 | 46 |
| 27 | 47 | 64 | 40 | 27 | 35 | 36 | 66 | 41 | 54 | 25 | 92 | 39 | 37 | 50 | 51 | 39 | 43 | 41 | 50 | 73 | [illegible] |
| 28 | 42 | 61 | 37 | 34 | 34 | 41 | 65 | 40 | 59 | 32 | 87 | 46 | 42 | 51 | 50 | 42 | 46 | 36 | 55 | 61 | 43 |
| 29 | 71 | 81 | 68 | 71 | 71 | 78 | 88 | 81 | 84 | 75 | 50 | 81 | 69 | 82 | 77 | 77 | 69 | 69 | 54 | 47 | [illegible] |
| 30 | 53 | 67 | 51 | 45 | 45 | 47 | 67 | 56 | 60 | 46 | 91 | 48 | 54 | 51 | [illegible] | 50 | 34 | 52 | 59 | 67 | [illegible] |
| 31 | 58 | 67 | 55 | 48 | 46 | 53 | 60 | 54 | 63 | 45 | 91 | 54 | 55 | 62 | 56 | 52 | 54 | 48 | 55 | 72 | [illegible] |

| | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|
| 24 | 50 | 83 | | | | | | | |
| 25 | 52 | 53 | 67 | | | | | | |
| 26 | 42 | 77 | 32 | 66 | | | | | |
| 27 | 25 | 74 | 45 | 63 | 43 | | | | |
| 28 | 30 | 77 | 48 | 66 | 44 | 41 | | | |
| 29 | 55 | 42 | 77 | 55 | 71 | [illegible] | 63 | | |
| 30 | 54 | 83 | 54 | 70 | 57 | 43 | 42 | 77 | |
| 31 | 42 | [illegible] | [illegible] | 64 | 48 | 35 | 46 | 61 | [illegible] |
| 32 | 71 | 56 | [illegible] | 75 | [illegible] | 64 | 67 | [illegible] | |

T[illegible]. Distanzen zwischen den Aufteilungen des 3. Datensatzes berechnet nach PAIRBONDS.

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | | | | |
| 2 | 1.529 | | | | | | | | | | | | | | | | | | | | |
| 3 | 1.331 | 1.322 | | | | | | | | | | | | | | | | | | | |
| 4 | 0.965 | 1.656 | 1.079 | | | | | | | | | | | | | | | | | | |
| 5 | 1.048 | 1.573 | 1.117 | 0.966 | | | | | | | | | | | | | | | | | |
| 6 | 1.124 | 1.934 | 1.685 | 0.968 | 1.014 | | | | | | | | | | | | | | | | |
| 7 | 1.137 | 1.434 | 1.307 | 1.157 | 1.053 | 1.344 | | | | | | | | | | | | | | | |
| 8 | 1.331 | 1.604 | 1.456 | 1.458 | 1.596 | 1.874 | 1.162 | | | | | | | | | | | | | | |
| 9 | 1.640 | 1.466 | 1.106 | 1.203 | 1.040 | 1.594 | 1.079 | 1.759 | | | | | | | | | | | | | |
| 10 | 1.593 | 1.691 | 1.708 | 1.430 | 1.844 | 1.554 | 1.755 | 1.137 | 1.669 | | | | | | | | | | | | |
| 11 | 1.595 | 1.622 | 2.086 | 2.259 | 1.993 | 2.206 | 1.108 | 1.840 | 1.868 | 2.418 | | | | | | | | | | | |
| 12 | 0.833 | 1.703 | 1.514 | 0.886 | 1.086 | 1.516 | 1.516 | 1.631 | 1.783 | 1.762 | 2.485 | | | | | | | | | | |
| 13 | 1.343 | 1.693 | 1.553 | 1.089 | 0.924 | 1.187 | 1.143 | 1.716 | 1.237 | 1.605 | 1.735 | 1.548 | | | | | | | | | |
| 14 | 2.015 | 1.824 | 1.782 | 2.085 | 1.609 | 2.138 | 2.105 | 2.055 | 1.945 | 2.006 | 2.333 | 1.995 | 1.959 | | | | | | | | |
| 15 | 1.328 | 1.676 | 1.313 | 1.360 | 1.325 | 1.789 | 1.233 | 1.234 | 1.171 | 1.827 | 1.912 | 1.316 | 1.517 | 1.443 | | | | | | | |
| 16 | 1.367 | 1.741 | 1.254 | 1.251 | 1.461 | 1.859 | 1.165 | 1.789 | 0.967 | 1.932 | 2.134 | 1.480 | 1.552 | 1.875 | 1.347 | | | | | | |
| 17 | 1.718 | 1.595 | 1.355 | 1.779 | 1.750 | 1.867 | 1.630 | 1.528 | 2.089 | 1.358 | 2.134 | 1.040 | 1.792 | 1.999 | 1.555 | 1.629 | | | | | |
| 18 | 1.506 | 1.945 | 1.110 | 1.743 | 1.516 | 2.031 | 1.590 | 1.624 | 1.536 | 1.616 | 2.346 | 1.860 | 1.655 | 1.809 | 1.421 | 1.407 | 1.528 | | | | |
| 19 | 1.702 | 1.257 | 0.799 | 1.588 | 1.746 | 2.046 | 1.720 | 1.834 | 1.715 | 1.644 | 2.203 | 1.724 | 1.977 | 1.132 | 1.481 | 1.598 | 1.255 | 1.320 | | | |
| 20 | 1.200 | 1.850 | 1.429 | 1.395 | 1.373 | 1.511 | 2.079 | 2.040 | 2.123 | 1.865 | 2.336 | 1.431 | 2.146 | 1.924 | 1.781 | 2.036 | 1.706 | 1.501 | 1.440 | | |
| 21 | 1.557 | 2.293 | 1.783 | 1.872 | 1.856 | 1.873 | 2.025 | 1.804 | 2.278 | 1.889 | 2.790 | 1.590 | 2.014 | 2.128 | 1.768 | 1.984 | 1.718 | 1.678 | 1.375 | 1.031 | |
| 22 | 1.003 | 1.742 | 1.442 | 1.251 | 1.836 | 1.256 | 1.696 | 1.473 | 2.054 | 1.195 | 2.464 | 1.431 | 1.904 | 2.045 | 1.648 | 1.723 | 1.344 | 1.512 | 1.579 | 1.010 | 1.565 |
| 23 | 1.881 | 2.102 | 1.603 | 2.082 | 1.974 | 2.010 | 1.917 | 1.958 | 2.276 | 1.758 | 1.947 | 1.876 | 2.030 | 1.947 | 2.005 | 1.864 | 1.474 | 1.535 | 1.544 | 1.627 | 1.559 |
| 24 | 1.292 | 1.232 | 0.802 | 1.229 | 1.327 | 1.941 | 1.174 | 1.381 | 0.714 | 1.617 | 1.830 | 1.165 | 1.467 | 1.651 | 1.124 | 0.936 | 1.554 | 1.182 | 1.112 | 1.838 | 1.97[illegible] |
| 25 | 1.552 | 1.673 | 1.619 | 1.542 | 1.556 | 1.532 | 1.307 | 1.519 | 1.800 | 1.845 | 1.336 | 1.593 | 1.529 | 1.928 | 1.527 | 1.538 | 1.525 | 1.690 | 1.892 | 1.830 | 1.740 |
| 26 | 1.321 | 1.450 | 1.077 | 1.657 | 1.580 | 1.833 | 1.863 | 1.840 | 1.635 | 1.561 | 2.456 | 1.523 | 1.777 | 1.615 | 1.461 | 1.566 | 1.365 | 0.950 | 0.507 | 1.1[illegible]4 | 1.543 |
| 27 | 1.523 | 1.991 | 1.723 | 1.395 | 1.983 | 1.640 | 2.087 | 1.684 | 1.820 | 1.383 | 2.847 | 1.258 | 1.827 | 1.970 | 1.677 | 1.786 | 1.534 | 1.646 | 1.470 | 1.351 | 1.713 |
| 28 | 1.040 | 1.751 | 1.426 | 1.451 | 1.477 | 1.932 | 1.558 | 1.394 | 1.865 | 1.331 | 2.547 | 1.419 | 1.971 | 2.033 | 1.599 | 1.722 | 1.519 | 1.395 | 1.625 | 1.010 | 1.065 |
| 29 | 1.442 | 1.876 | 1.512 | 1.820 | 1.811 | 1.936 | 1.912 | 2.052 | 2.165 | 2.007 | 1.903 | 1.792 | 1.819 | 1.744 | 1.924 | 1.969 | 1.495 | 1.514 | 1.363 | 0.868 | 0.913 |
| 30 | 1.700 | 1.893 | 2.017 | 1.885 | 1.779 | 1.911 | 1.973 | 1.923 | 2.171 | 2.068 | 2.574 | 1.647 | 2.202 | 1.704 | 1.916 | 1.934 | 1.894 | 2.029 | 1.655 | 1.367 | 1.699 |
| 31 | 1.742 | 1.933 | 2.078 | 1.963 | 1.946 | 2.065 | 1.892 | 2.034 | 2.179 | 2.112 | 2.617 | 1.692 | 2.391 | 1.831 | 1.738 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |

| | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | 1.537 | | | | | | | | | | |
| 24 | 1.636 | 1.770 | | | | | | | | | |
| 25 | 1.421 | 1.076 | 1.481 | | | | | | | | |
| 26 | 1.195 | 1.598 | 1.114 | 1.734 | | | | | | | |
| 27 | 1.060 | 1.441 | 1.586 | 1.740 | 1.525 | | | | | | |
| 28 | 1.102 | 1.635 | 1.570 | 1.576 | 1.693 | 1.702 | | | | | |
| 29 | 1.198 | 0.859 | 1.725 | 1.163 | 1.466 | 1.525 | 1.352 | | | | |
| 30 | 1.995 | 1.893 | 1.841 | 1.932 | 1.978 | 1.912 | 1.796 | 1.818 | | | |
| 31 | 1.617 | 1.496 | 1.771 | 1.785 | 1.529 | 1.604 | 1.997 | 1.446 | 1.469 | | |
| 32 | 1.845 | 1.190 | 1.610 | 1.961 | 1.518 | 1.577 | 1.839 | 1.324 | 1.453 | 1.359 | |

Tab. E : Distanzen zwischen den Aufteilungen des

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2.072 | | | | | | | | | | |
| 3 | 2.026 | 1.853 | | | | | | | | | |
| 4 | 1.323 | 2.435 | 1.628 | | | | | | | | |
| 5 | 1.523 | 2.405 | 1.625 | 1.151 | | | | | | | |
| 6 | 1.741 | 3.053 | 2.637 | 1.618 | 1.694 | | | | | | |
| 7 | 1.979 | 2.181 | 2.204 | 1.898 | 1.747 | 2.355 | | | | | |
| 8 | 1.602 | 2.255 | 2.173 | 1.680 | 1.971 | 2.790 | 1.730 | | | | |
| 9 | 1.937 | 2.139 | 2.044 | 1.740 | 1.740 | 2.483 | 1.840 | 1.970 | | | |
| 10 | 1.617 | 2.320 | 2.179 | 1.637 | 2.112 | 2.169 | 2.547 | 1.457 | 1.864 | | |
| 11 | 1.364 | 1.366 | 2.078 | 2.257 | 1.860 | 2.305 | 1.219 | 1.676 | 1.986 | 2.611 | |
| 12 | 1.302 | 2.526 | 2.173 | 1.145 | 1.420 | 2.239 | 2.006 | 2.041 | 2.134 | 2.106 | 2.480 |
| 13 | 1.600 | 2.280 | 2.186 | 1.518 | 1.394 | 1.776 | 1.828 | 2.139 | 1.858 | 1.718 | 1.654 |
| 14 | 2.492 | 2.375 | 2.473 | 2.931 | 2.406 | 2.939 | 2.867 | 2.826 | 2.746 | 2.401 | 2.466 |
| 15 | 1.729 | 2.461 | 1.825 | 2.043 | 1.994 | 2.976 | 1.677 | 1.988 | 1.822 | 2.443 | 1.903 |
| 16 | 1.8 4 | 2.661 | 1.786 | 1.692 | 2.043 | 2.851 | 1.953 | 2.512 | 1.507 | 2.443 | 2.054 |
| 17 | 2.4 9 | 2.710 | 1.993 | 2.335 | 2.411 | 2.744 | 2.196 | 2.107 | 2.675 | 1.935 | 2.346 |
| 18 | 1.815 | 2.603 | 1.701 | 2.044 | 2.044 | 3.052 | 2.419 | 1.779 | 2.184 | 1.719 | 2.408 |
| 19 | 2.479 | 1.8 2 | 1.290 | 2.233 | 2.433 | 3.241 | 2.694 | 2.453 | 2.421 | 2.308 | 2.243 |
| 20 | 2.0 4 | 2.173 | 1.861 | 2.204 | 2.253 | 2.576 | 2.779 | 2.749 | 2.868 | 2.604 | 2.213 |
| 21 | 2.303 | 3.226 | 2.448 | 2.361 | 2.456 | 2.713 | 3.042 | 2.201 | 2.970 | 2.105 | 2.916 |
| 22 | 1.375 | 2.459 | 1.917 | 1.748 | 2.450 | 2.032 | 2.530 | 2.119 | 2.639 | 1.574 | 2.539 |
| 23 | 2.416 | 2.575 | 2.473 | 2.979 | 2.855 | 2.912 | 2.515 | 2.751 | 2.334 | 2.730 | 1.950 |
| 24 | 1.621 | 1.734 | 1.753 | 1.440 | 1.640 | 2.809 | 1.776 | 1.546 | 1.100 | 1.840 | 1.722 |
| 25 | 2.051 | 1.527 | 2.027 | 2.004 | 2.079 | 2.438 | 2.114 | 2.063 | 2.458 | 2.555 | 1.352 |
| 26 | 2.077 | 2.000 | 1.163 | 2.196 | 2.030 | 2.639 | 2.642 | 2.526 | 2.295 | 1.906 | 2.517 |
| 27 | 2.245 | [illegible] | 2.551 | 1.743 | 2.494 | 2.351 | 2.955 | 2.239 | 2.483 | 1.618 | 3.279 |
| 28 | 1.379 | 2.675 | 2.048 | 1.981 | 2.056 | 2.513 | 2.642 | 1.925 | 2.645 | 1.981 | 2.792 |
| 29 | [illegible] | [illegible] | 2.171 | 2.428 | 2.504 | 2.837 | 2.614 | 2.649 | 2.607 | 2.779 | 1.874 |
| 30 | 2.651 | [illegible] | 2.877 | 2.619 | [illegible] | 2.612 | [illegible] | 2.900 | 3.144 | 2.755 | 2.815 |
| 31 | [illegible] | [illegible] | [illegible] | 2.679 | 2.679 | 3.012 | [illegible] | 2.749 | [illegible] | 2.579 | [illegible] |
| 32 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 3.174 | [illegible] | 2.737 | 2.361 | 2.942 | [illegible] |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 1.788 | | | | | | | | | |
| 14 | 2.551 | 2.488 | | | | | | | | |
| 15 | 2.037 | 2.200 | 2.712 | | | | | | | |
| 16 | 1.912 | 2.200 | 2.612 | 1.849 | | | | | | |
| 17 | 2.195 | 2.293 | 2.780 | 2.607 | 2.293 | | | | | |
| 18 | 2.228 | 2.001 | 2.203 | 2.050 | 1.925 | 2.044 | | | | |
| 19 | 2.216 | 2.790 | 1.992 | 2.265 | 2.390 | 2.033 | 1.640 | | | |
| 20 | 2.312 | 3.010 | 2.563 | 2.810 | 2.935 | 2.083 | 1.764 | 1.524 | | |
| 21 | 2.040 | 2.538 | 2.625 | 2.587 | 2.662 | 2.069 | 1.863 | 1.767 | 1.734 | |
| 22 | 2.168 | 2.456 | 2.895 | 2.095 | 2.381 | 1.539 | 1.896 | 2.321 | 1.695 | 1.918 |
| 23 | 2.449 | 2.812 | 2.614 | 2.776 | 2.737 | 1.819 | 1.976 | 1.841 | 1.698 | 1.425 |
| 24 | 1.424 | 1.683 | 2.160 | 1.758 | 1.358 | 2.326 | 1.408 | 1.321 | 2.463 | 2.420 |
| 25 | 2.424 | 2.161 | 2.639 | 2.200 | 2.161 | 1.919 | 2.125 | 2.426 | 2.122 | 2.039 |
| 26 | 2.090 | 2.112 | 1.914 | 1.902 | 1.985 | 1.831 | 1.313 | 0.888 | 1.795 | 1.764 |
| 27 | 1.963 | 2.376 | 2.939 | 2.576 | 2.576 | 2.144 | 2.301 | 2.365 | 2.661 | 2.313 |
| 28 | 2.325 | 2.413 | 2.826 | 2.387 | 2.662 | 2.230 | 1.853 | 2.652 | 1.823 | 1.649 |
| 29 | 2.573 | 2.425 | 2.653 | 2.535 | 2.786 | 2.144 | 2.074 | 1.561 | 1.257 | 1.250 |
| 30 | 2.424 | 3.112 | 2.449 | 2.086 | 2.837 | 2.529 | 2.656 | 2.675 | 2.286 | 2.574 |
| 31 | 2.549 | 2.837 | 2.449 | 2.535 | 2.961 | 2.454 | 2.211 | 2.324 | 2.400 | 2.299 |
| 32 | 2.111 | 3.074 | 2.476 | 2.448 | 2.773 | 2.406 | 2.739 | 1.838 | 1.534 | 1.771 |

| | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 2.339 | 2.446 | | | | | | | | | |
| 25 | 1.732 | 1.373 | 1.958 | | | | | | | | |
| 26 | 1.844 | 2.314 | 1.395 | 2.224 | | | | | | | |
| 27 | 1.281 | 2.312 | 2.258 | 2.537 | 2.163 | | | | | | |
| 28 | 1.594 | 2.550 | 2.345 | 2.299 | 2.250 | 2.113 | | | | | |
| 29 | 1.595 | 1.148 | 2.268 | 1.537 | 2.124 | 2.236 | 1.972 | | | | |
| 30 | 2.968 | 2.773 | 2.444 | 2.798 | 2.549 | 2.612 | 2.374 | 2.747 | | | |
| 31 | 2.217 | 2.137 | 2.368 | 2.561 | 2.198 | 2.061 | 2.699 | 2.061 | 2.047 | | |
| 32 | 2.730 | 1.633 | 1.996 | 2.546 | 2.111 | 2.574 | 2.661 | 1.934 | 1.911 | 1.674 | |

Tab. L : Distanzen zwischen den Aufteilungen des
nach INFOTWO

Tab. M : Distanzen zwischen den Aufteilungen des 3.Datensatzes nach INFOTHRY.

Abb. A 1 Computer-Ausdruck eines Baumgraphen
(Aufteilungen des 3.Datensatzes, ROSSBACH/Ward)

## Lebenslauf

Am 19. Juli 1943 wurde ich als Sohn des kaufmännischen Angestellten Heinz Roßbach und seiner Ehefrau Hildegard, geb. Friedlein, in Hamburg geboren. Ich besuchte dort die Volksschule von 1950 - 1954 und anschließend das 'Gymnasium für Jungen in Eimsbüttel' auf dem mathematisch-naturwissenschaftlichen Zweig, wo ich im Frühjahr 1963 das Abitur erlangte.

Im Sommersemester desselben Jahres nahm ich in Hamburg das Studium der Naturwissenschaften im Fach Biologie auf, um dann aber schon im Wintersemester auf die Psychologie überzuwechseln. Im Mai 1967 legte ich die Diplom-Vor- und im Juni 1969 die Diplom-Hauptprüfung in Psychologie ab.

Von April bis September 1970 war ich als wissenschaftlicher Angestellter bei Prof. Dr. P.R. Hofstätter am Psychologischen Institut der Universität Hamburg tätig. Anschließend im Rahmen eines von der 'Wissenschaftlichen Gesellschaft des Saarlandes' geförderte Forschungsvorhabens über Taxonomische Methoden bei Herrn Prof. Dr. W.H. Tack an der Universität des Saarlandes, Saarbrücken. Seit April dieses Jahres bin ich wissenschaftlicher Mitarbeiter in einem Projekt der Deutschen Forschungsgemeinschaft über Empirische Kriminologie unter Leitung von Herrn Prof. Dr. A. Baratta.